Economics Letters 156 (2017) 138–141

Contents lists available at ScienceDirect

Economics Letters

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

Quasi-generalized least squares regression estimation with spatial data☆

Cuicui Lu [a], Jeffrey M. Wooldridge [b,*]

[a] Department of Economics, Nanjing University, China
[b] Department of Economics, Michigan State University, USA

ARTICLE INFO

Article history:
Received 27 March 2017
Accepted 5 April 2017
Available online 28 April 2017

JEL classification:
C13

Keywords:
Quasi-GLS
Spatial correlation
Covariance tapering
Spatial HAC estimator

ABSTRACT

We use a particular quasi-generalized least squares (QGLS) approach to study a linear regression model with spatially correlated error terms. The QGLS estimator is consistent, asymptotically normal, computationally easier than GLS, and it appears to not lose much efficiency. A variance–covariance estimator for QGLS, which is robust to heteroskedasticity, spatial correlation and general variance–covariance misspecification is provided.

1. Introduction

Economists often use proximity to measure interactions among agents. Even if one is not interested in interactions, one often needs to account for spatial dependence in cross-sectional data when conducting inference.

There are two popular approaches to accounting for spatial correlation among the errors in linear models. First, one can apply ordinary least squares (OLS) for estimation and then obtain standard errors and test statistics that are robust to fairly general forms of spatial correlation. Though computationally simple, OLS may be quite inefficient. The second approach is to apply feasible generalized least squares (FGLS) in an attempt to improve efficiency over OLS.

There are some potential drawbacks with FGLS. First, with a sample of size N, we cannot estimate an $N \times N$ variance–covariance matrix without imposing restrictions. Second, the calculation of the inverse of a huge variance–covariance matrix generally needs substantial computer memory and can run slowly. Third, most FGLS approaches conduct inference as if the fully variance–covariance matrix is correctly specified, which can be very misleading in practice.

In this paper we propose a middle ground between OLS and a fully specified FGLS analysis. Our approach gains back much of the efficiency lost by using OLS while being computationally fairly simple. The method we propose, quasi-generalized least squares (quasi-GLS or QGLS), uses observations of nearest neighbors in a GLS-type analysis. We use the modifier "quasi" because we understand that, by ignoring units that are not in the groups that we form, we are unlikely conducting full GLS. Our QGLS estimator intentionally sets the vast majority of covariances to zero. Since the correlations within nearby units accounts for most of the dependence in the data, it is possible to get an estimator that is close to the full FGLS estimator in terms of asymptotic efficiency. In addition, it is straightforward to obtain inference robust to general spatial correlation patterns.

2. A linear model

Spatial data can be analyzed using cross section or panel data. For example, Baltagi and Pirotte (2011) study estimation in the context of unobserved effects panel data models with spatial correlation, where they assume, like other authors, a correctly specified variance–covariance matrix. Because we are proposing a new approach that does not even assume correct specification of

☆ This paper is supported by the National Natural Science Foundation of China, No.71601094.

* Corresponding author.
E-mail addresses: lucuicui@nju.edu.cn (C. Lu), wooldri1@msu.edu (J.M. Wooldridge).

http://dx.doi.org/10.1016/j.econlet.2017.04.004
0165-1765/

the spatial variances and covariances, we restrict attention to the cross section case.

Our notation is adopted from Jenish and Prucha (2009). $\mathcal{S}$ is the space the population resides. Let $\{(\mathbf{x}_i, y_i), i = 1, 2, \ldots, N\}$ denote the data sampled at location $s_i \in \mathcal{S}$. Let $\{u_i, i = 1, 2, \ldots, N\}$ denote the underlying error process. Let d_{ij} be the distance between location s_i and s_j and denote the collection of distances by $\mathbf{D} \equiv \{d_{ij}, i, j = 1, 2, \ldots, N\}$. The model is

$$y_i = \mathbf{x}_i\beta + u_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

where $\mathbf{x}_i$ is a $1 \times K$ vector of regressors with $x_{i1} = 1$ and $\beta \equiv (\beta_1, \beta_2, \ldots, \beta_K)'$ is a $K \times 1$ unknown vector of parameters. In matrix form,

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

where $\mathbf{y} = (y_1, y_2, \ldots, y_N)'$, $\mathbf{X}$ is an $N \times K$ matrix, with the ith row equal to $\mathbf{x}_i$, and $\mathbf{u} = (u_1, u_2, \ldots, u_N)'$. We allow $\mathbf{u}$ to exhibit spatial correlation:

$$\mathrm{Var}(\mathbf{u}|\mathbf{X}, \mathbf{D}) = \mathbf{\Omega}(\mathbf{D}, \lambda), \tag{3}$$

where λ is a vector of variance–covariance parameters. For example, we might propose

$$\mathbf{\Omega}_{ii} = \sigma^2, \tag{4}$$

and

$$\mathbf{\Omega}_{ij} = \sigma^2 c(d_{ij}, \rho), \quad i \neq j, \tag{5}$$

where ρ is a spatial correlation parameter and $c(\cdot)$ is a correlation function that decreases in d_{ij}.

3. Estimation

We want to show proof of concept as opposed to deriving new asymptotic theory. The asymptotic results in Jenish and Prucha (2009) can be applied immediately to the OLS estimator, and they are easily modified for the quasi-FGLS estimator. As in Jenish and Prucha (2009), we are thinking of increasing domain asymptotics.

3.1. OLS estimator

In addition to being of interest in its own right, OLS provides a first-stage estimation for QGLS. Given the data vector $\mathbf{y}$ and data matrix $\mathbf{X}$, the OLS estimator is

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{6}$$

Under the assumptions of Theorem 1 in Jenish and Prucha (2009), $\hat{\beta}_{OLS}$ is consistent and asymptotically normal:

$$\sqrt{N}\left(\hat{\beta}_{OLS} - \beta\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\right), \tag{7}$$

where the limits

$$\mathbf{A} = \lim_{\mathrm{N}\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}\left(\mathbf{x}_i'\mathbf{x}_i\right) \tag{8}$$

$$\mathbf{B} = \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i'u_i\right) \tag{9}$$

are assumed to exist, with $\mathbf{A}$ being nonsingular. Estimation of $\mathbf{A}$ is straightforward using the sample average, and estimation of $\mathbf{B}$ can be done as in Conley (1999) or Kelejian and Prucha (2007).

3.2. Quasi-GLS estimator

Our quasi-GLS approach starts by dividing the spatial data into groups based on "closeness," so that correlation within groups is relatively large. Admittedly, the group choices are somewhat arbitrary, but that choice does not affect consistency or asymptotically normality under standard assumptions. If we assume a correct structure for the variance–covariance matrix, the most efficient estimator is full FGLS. However, such an estimator is computationally intensive, and it need not be efficient if we allow a misspecified variance–covariance structure.

For our approach, we have in mind relatively few units per group: at least two and perhaps up to a handful. The exact choice of groups is left to future research. Once we have chosen the groups, we can think of a system of equations

$$\mathbf{y}_g = \mathbf{X}_g\beta + \mathbf{u}_g, \quad g = 1, 2, \ldots, G, \tag{10}$$

where each group has the same size L (although this is not required). Therefore, $\mathbf{y}_g$ and $\mathbf{u}_g$ are $L \times 1$ and $\mathbf{X}_g$ is $L \times K$. Let $\mathbf{\Lambda}_g \equiv \mathrm{Var}\left(\mathbf{u}_g|\mathbf{X}_g, \mathbf{D}_g\right)$ be the variance–covariance matrix for group g, where $\mathbf{D}_g$ is the lattice where the observations in group g reside. Because we are treating the distances as nonrandom, we are free to group observations by distance without causing inconsistency. Let $\mathbf{\Omega}_N$ be the true $N \times N$ variance–covariance matrix and $\mathbf{\Lambda}_N$ be the block-diagonal matrix that only contain within group variances and covariances.

Suppose $\mathbf{\Lambda}_g$ is dependent on the parameter vector λ and $\hat{\lambda}$ is an estimator for λ and $\hat{\mathbf{\Lambda}}_g$ the corresponding estimator of $\mathbf{\Lambda}_g$. The feasible QGLS (FQGLS) estimator can be written as

$$\hat{\beta}_{FQGLS} = \left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\hat{\mathbf{\Lambda}}_g^{-1}\mathbf{y}_g\right). \tag{11}$$

As with the OLS estimator, we do not provide careful regularity conditions, as they are standard in the spatial econometrics literature. One of the differences with pooled OLS is that we must assume the explanatory variables are strictly exogenous in the sense that

$$\mathrm{E}(u_i|\mathbf{x}_1, \mathbf{x}_2, \ldots) = 0, \quad i = 1, 2, \ldots, \tag{12}$$

which implies $\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\right) = \mathbf{0}$ for all g. In addition to spatial mixing conditions we assume that

$$\mathbf{Q} = \lim_{G\to\infty}\frac{1}{G}\sum_{g=1}^{G}\mathrm{E}\left(\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{X}_g\right) \tag{13}$$

and

$$\mathbf{S} = \lim_{G\to\infty} E\left(\frac{1}{G}\sum_{g=1}^{G}\sum_{h=1}^{G}\mathbf{X}_g'\mathbf{\Lambda}_g^{-1}\mathbf{u}_g\mathbf{u}_h'\mathbf{\Lambda}_h^{-1}\mathbf{X}_h\right) \tag{14}$$

exist and $\mathbf{Q}$ has rank K. We can apply the results in Jenish and Prucha (2009) because we are grouping "nearby" observations, and so the groups will form a spatially mixing sequence and generally satisfy Theorems 1 and 3 in Jenish and Prucha (2009). Therefore, it is generally true that

$$\sqrt{G}\left(\hat{\beta}_{FQGLS} - \beta\right) \to^{d} \mathrm{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1}\right). \tag{15}$$

The result in (15) allows nonnormal errors and also three kinds of misspecification in the variance–covariance matrix. One intentional misspecification is that we ignore correlations among observations not in the same group. Second, we allow for misspecification of the structure of Λ_g, which holds if we do not have the original structure of Ω_N correctly specified. Third, even if we correctly specify Λ_g we allow the estimator of λ to be inconsistent. The work here differs from Andrews and Guggenberger (2012) in that we allow misspecification in both the heteroskedasticity and spatial correlation structures.