Journal of Statistical Planning and Inference 140 (2010) 2470–2477

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A working estimating equation for spatial count data

Pei-Sheng Lin [a,b,*]

[a] Division of Biostatistics and Bioinformatics, National Health Research Institute, 35 Keyan Road, Miaoli 350, Taiwan
[b] Department of Mathematics, National Chung Cheng University, Taiwan

ARTICLE INFO

Article history:
Received 29 October 2009
Received in revised form
21 February 2010
Accepted 23 February 2010
Available online 2 March 2010

Keywords:
Estimating equation
Parameter-driven model
Quasi-likelihood function
Spatial count data

ABSTRACT

This paper proposes a working estimating equation which is computationally easy to use for spatial count data. The proposed estimating equation is a modification of quasi-likelihood estimating equations without the need of correctly specifying the covariance matrix. Under some regularity conditions, we show that the proposed estimator has consistency and asymptotic normality. A simulation comparison also indicates that the proposed method has competitive performance in dealing with over-dispersion data from a parameter-driven model.

1. Introduction

Spatially correlated count data arise in a variety of settings. However, due to complicated dependence between observations, using a full-likelihood approach to carry out the estimation usually involves computational intensity, or the likelihood function is even untractable (McCullagh, 1991). One possibility to address this is to use an estimating equation that only requires specification of the first two moments.

There are various authors to propose estimating equations for correlated data. For binary responses, Heagerty and Lele (1998) proposed an estimating equation which only depends on their marginal means and pairwise covariances. Oman et al. (2007) further generalized the work of Heagerty and Lele by constructing an estimating equation with approximation to the pairwise covariances. Heagerty and Lumley (2000) used the concept of quasi-likelihood (QL) function to model an estimating equation for marginal mean of count data. For a parameter-driven model with correlation being introduced through a latent process, Zeger (1988) proposed a two-stage estimating equation for time series models. McShane et al. (1997) extended Zeger's approach to a spatial setting by using generalized estimating equations (GEEs, Liang and Zeger, 1986) to allow for marginal inferences on mean structure parameters of a spatial count process. Nevertheless, the estimating equation proposed by McShane et al. was built based on the fully specified covariance structure of the latent process.

In reality, we may only have incomplete information about the covariance structure of responses, or may not even know whether a latent process exists behind the observations. In this paper, we propose a working estimating equation generalized from the QL function to estimate unknown parameters for spatial count data. In the proposed estimating equation, only partial information about the covariance needs to be specified. We show that, under some regularity conditions, the proposed estimator is consistent and asymptotically normal in Section 2. In Section 3, we describe the parameter-driven model and another related estimation method proposed by McShane et al. (1997). A series of

* Correspondence address: Division of Biostatistics and Bioinformatics, National Health Research Institute, 35 Keyan Road, Miaoli 350, Taiwan.
E-mail addresses: pslin@math.ccu.edu.tw, pslin@nhr.org.tw.

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.02.019

simulations based on parameter-driven models is then conducted to study the performance of the proposed estimator against mis-specification of covariances in Section 4. Some discussions are in Section 5.

2. Quasi-likelihood working estimating equations

The QL concept was originally proposed by Wedderburn (1974) for an extension of generalized least squares. In the QL estimation, we only consider the marginal mean and covariance structure of responses. Let $\boldsymbol{Y}=\{Y(\boldsymbol{s}_1),\ldots,Y(\boldsymbol{s}_n)\}$ be observations from a random field $Y(\boldsymbol{s})$, $\boldsymbol{s}\in R^d$, where $d\geq 2$. At location $\boldsymbol{s}_i$, we assume that the mean response of $Y(\boldsymbol{s}_i)$ is associated with an explanatory vector $\boldsymbol{x}(\boldsymbol{s}_i)$ through a link function $g\{\theta(\boldsymbol{s}_i)\}=\boldsymbol{x}^T(\boldsymbol{s}_i)\boldsymbol{\beta}$, where $\theta(\boldsymbol{s}_i)=E\{Y(\boldsymbol{s}_i)\}$ and $\boldsymbol{\beta}$ is a parameter vector of interest. Without ambiguity of notation, we also use Y_i, $\boldsymbol{x}_i$, and θ_i to represent $Y(\boldsymbol{s}_i)$, $\boldsymbol{x}(\boldsymbol{s}_i)$, and $\theta(\boldsymbol{s}_i)$, respectively. Moreover, suppose that there exists a smooth variance function $V(\cdot)$ such that $\mathrm{Var}(Y_i|\boldsymbol{x}_i)=V\{g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})\}$, which represents the variance is a function of the mean only. A marginal model satisfying the above assumptions is called a QL type regression model.

Let $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)^T$ and $\boldsymbol{V}$ denote the mean vector and covariance matrix of the responses, respectively. Also, we define $\boldsymbol{D}=\partial\boldsymbol{\theta}/\partial\boldsymbol{\beta}$ to represent the Jacobian matrix between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Using a matrix $\boldsymbol{D}^T\boldsymbol{V}^{-1}$ to project the centered responses $\boldsymbol{Y}-\boldsymbol{\theta}$ into the spanning space formed by $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$, the QL estimating equation

$$\boldsymbol{U}(\boldsymbol{\beta})=\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\theta})=\boldsymbol{0} \tag{2.1}$$

is constructed to obtain possibly maximum information from data (Li, 1996). The solution of (2.1) is called a QL estimator $\hat{\boldsymbol{\beta}}_{QLE}$. When observations are independently drawn from an exponential family, McCullagh (1983) showed that $\hat{\boldsymbol{\beta}}_{QLE}$ is asymptotically optimal among all linear unbiased estimators.

When data are correlated in multi-dimensional spaces, the literature concerning asymptotic properties of the QL estimator is relatively limited. We next introduce a central limit theorem developed by Lin (2008) for dependent random fields. (Other central limit theorems related with mixing coefficients can be found in Guyon, 1995.) Let Λ denote a lattice subset of locations, and $|\Lambda|$ the number of elements in Λ. We define a mixing coefficient by

$$\rho_{k,l;p}(m)=\sup\left[\left|\mathrm{cov}\left\{\prod_{\boldsymbol{s}_i\in\Lambda_1}Y(\boldsymbol{s}_i),\prod_{\boldsymbol{s}_j\in\Lambda_2}Y(\boldsymbol{s}_j)\right\}\right|:|\Lambda_1|\leq k,|\Lambda_2|\leq l,d_p(\Lambda_1,\Lambda_2)\geq m\right],$$

where $d_p(\Lambda_1,\Lambda_2)=\inf\{\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_p:\boldsymbol{s}_i\in\Lambda_i\}$, $\|\boldsymbol{s}_1-\boldsymbol{s}_2\|_p$ is an L_p norm between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. The coefficient $\rho_{k,l;p}(m)$ is defined to control correlation coefficients for products of random variables. Let $S_n=\sum_{\boldsymbol{s}\in\Lambda_n}\{Y(\boldsymbol{s})-\mu(\boldsymbol{s})\}$ and $\tau_n^2=\mathrm{var}(S_n)$, where Λ_n is a strictly increasing subsequence of lattice sets. Under appropriate mixing conditions, it can be shown that S_n converges in distribution to a normal distribution.

Lemma 1. *Suppose that a lattice random field satisfies*

(i) $\rho_{1,1;p}(m)=\mathrm{O}(m^{-d-\varepsilon})$ *for some* $\varepsilon>0$, *and*
(ii) $\rho_{k,l;p}(m)=\mathrm{o}(m^{-d})$ *for* $k+l\leq 4$.

Then S_n/τ_n *converges in distribution to* N(0,1).

Lemma 1 holds for spatial processes with exponential or spherical correlations. A much broader class of processes for Theorem 1 can be seen in Lin (2008). Now, suppose that we only have incomplete information for the covariance structure of $\boldsymbol{Y}$. So, instead of using the exact covariance matrix $\boldsymbol{V}$ to construct a QL estimating equation, for spatial count data, we propose to use a working covariance matrix

$$(\boldsymbol{\Omega}_n)_{i,j}=\theta_i\theta_j\hat{\rho}_{\boldsymbol{\eta}}(h_{i,j}) \tag{2.2}$$

to construct an estimating equation

$$\boldsymbol{Q}_n(\boldsymbol{\beta})=\boldsymbol{D}^T\boldsymbol{\Omega}_n^{-1}(\boldsymbol{Y}-\boldsymbol{\theta})=\boldsymbol{0}, \tag{2.3}$$

where $h_{i,j}=\|\boldsymbol{s}_i-\boldsymbol{s}_j\|_p$ and $\hat{\rho}_{\boldsymbol{\eta}}(h)$ is an estimated positive definite correlation model with parameters $\boldsymbol{\eta}$. When $\boldsymbol{\Omega}_n=\boldsymbol{V}$, the proposed estimating equation is equivalent to the QL estimating equation. Note that the derivative matrix of $\boldsymbol{Q}_n(\boldsymbol{\beta})$ is symmetric, and hence $\boldsymbol{Q}_n(\boldsymbol{\beta})$ is a gradient vector of some quasi-likelihood functions (McCullagh and Nelder, 1989). The proposed estimating equation therefore satisfies the required conditions of existence, which usually is an obstacle to apply the QL estimating equation to dependent data (Li, 1996). Moreover, the following theorem states that the working estimating equation based on (2.2) and (2.3) could still have asymptotic normality.

Theorem 1. *Assume that the following conditions hold for an estimating equation* $\boldsymbol{Q}_n(\boldsymbol{\beta})$ *with a working covariance matrix* (2.2):

(A) *The covariates are finite. That is*, $\max_{i,j}|x_{i,j}|<\infty$.
(B) *The link function* $g(\theta)$ *is smooth over* θ.