Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Generalized estimating equations with stabilized working correlation structure

Q2 Yongchan Kwon[a], Young-Geun Choi[a], Taesung Park[a], Andreas Ziegler[b,c,d], Myunghee Cho Paik[a,*]

[a] *Seoul National University, Seoul, Republic of Korea*
[b] *Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Campus Lübeck, Germany*
[c] *Center for Clinical Trials, ZKS Lübeck, University of Lübeck, Germany*
[d] *School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa*

ARTICLE INFO

Article history:
Received 22 April 2016
Received in revised form 19 August 2016
Accepted 26 August 2016
Available online xxxx

Keywords:
Generalized estimating equations
Working correlation
Unstructured correlation matrix
Positive definiteness
Linear shrinkage

ABSTRACT

Generalized estimating equations (GEE) proposed by Liang and Zeger (1986) yield a consistent estimator for the regression parameter without correctly specifying the correlation structure of the repeatedly measured outcomes. It is well known that the efficiency of regression coefficient estimator increases with correctly specified working correlation and thus unstructured correlation could be a good candidate. However, lack of positive-definiteness of the estimated correlation matrix in unbalanced case causes practitioners to choose independent, autoregressive or exchangeable matrices as working correlation structure. Our goal is to broaden practical choices of working correlation structure to unstructured correlation matrix or any other matrices by proposing a GEE with a stabilized working correlation matrix via linear shrinkage method in which the minimum eigenvalue is forced to be bounded below by a small positive number. We show that the resulting regression estimator of GEE is asymptotically equivalent to that of the original GEE. Simulation studies show that the proposed modification can stabilize the variance of the GEE regression estimator with unstructured working correlation, and improve efficiency over popular choices of working correlation. Two real data examples are presented where the standard error of the regression coefficient estimator can be reduced using the proposed method.

1. Introduction

Generalized estimating equations (GEE) proposed by Liang and Zeger (1986) have been a popular analytic tool for correlated data. A consistent estimator for the regression parameter can be achieved without correctly specifying the correlation structure of the repeatedly measured outcomes. However, the efficiency of regression coefficient estimator increases if the working correlation matrix is close to the true one (Albert and McShane, 1995). Structured working correlations such as independent, autoregressive and exchangeable are available from built-in functions from software. These choices give a manageable number of parameters in the correlation matrix, and can be helpful when the sample size is small and the number of time points is large. To select a working correlation matrix from various choices, criteria such

[*] Corresponding author.
E-mail address: myungheechopaik@snu.ac.kr (M.C. Paik).

http://dx.doi.org/10.1016/j.csda.2016.08.016
0167-9473/

Please cite this article in press as: Kwon, Y., et al., Generalized estimating equations with stabilized working correlation structure. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.08.016

as the 'quasi-likelihood under the independence model criterion' (Pan, 2001) and the 'correlation information criterion' (Hin and Wang, 2009) have been proposed among others (Carey and Wang, 2011; Gosho et al., 2011; Zhou et al., 2012; Westgate, 2013, 2014). The unstructured working correlation matrix can correctly model the correlation structure and is available from built-in functions from software, but the number of unknown parameters increases as the number of time points. When the sample size is small relative to the number of time points, variability of many nuisance parameters in the unstructured correlation matrix affects the variance of the regression parameter estimators, and Westgate (2013) proposed a method to address this problem. However, when the maximum of numbers of time points is fixed, the asymptotic variance of the regression coefficient estimator is unaffected by the variance of the correlation estimator, and reducing the number of parameters does not lead to gain in asymptotic efficiency of the regression coefficient estimator. Misspecification of working correlation could not only lead to loss of efficiency, but more seriously, could lead to infeasibility of the GEE solutions (Qu et al., 2008; Wang and Carey, 2004). Despite these shortcomings, choosing aforementioned structured working correlation matrix guarantees the correlation matrix to be positive definite. The estimated unstructured correlation matrix sometimes fails to be positive definite due to varying numbers of subunits, in which case the GEE estimates are not defined. Even when the estimated unstructured matrix is positive definite, if the minimum eigenvalue is small, the coefficient estimate can be unstable and the standard error of regression parameter estimates can be large (Vens and Ziegler, 2012). If lack of positive definiteness can be solved, the unstructured working correlation matrix can be an attractive choice since it improves the asymptotic variance of the regression coefficient estimator.

Many researchers have worked on solving lack of positive-definiteness of the sample covariance matrix mainly by replacing the eigenvalues of sample covariance matrix by their linear or nonlinear transforms (Stein, 1956; Haff, 1991; Daniels and Kass, 1999, 2001; Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005; Ledoit and Wolf, 2012; Won et al., 2013; Lam, 2016). In a regression setting with longitudinal data, Daniels and Kass (2001) obtained stabilized regression coefficients estimators by placing a normally-distributed prior to the logarithm of the sample eigenvalues. This method requires that the eigenvalues of the sample covariance matrix are positive.

Our goal is to broaden practical choices of working correlation structure to unstructured correlation matrix by alleviating problems due to lack of positive definiteness. To achieve this goal we propose to modify working correlation matrix by linear shrinkage method proposed by Choi (2015). We show that the resulting regression estimator of GEE is asymptotically equivalent to that of the original GEE. Simulation studies show that the proposed modification has advantages in cases where the minimum eigenvalue of the estimated working correlation structure is small. Two real data examples are presented where the standard error of the regression coefficient estimator is reduced using the proposed method.

2. Basic notations

We denote the $n_i \times 1$ vector of the outcomes and the $n_i \times p$ matrix of covariates for the ith subject ($i = 1, \ldots, K$) by $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ and $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})^T$, respectively. We assume that the first two moments of y_{ij} are given by

$$E(y_{ij} \mid x_{ij}) = \mu_{ij} = g(\eta_{ij}) = g(x_{ij}^T\boldsymbol{\beta}), \quad \text{and} \quad Var(y_{ij} \mid x_{ij}) = \phi a(\mu_{ij}),$$

where $\boldsymbol{\beta}$ is a $p \times 1$ regression parameter, and $g^{-1}(\cdot)$ is a link function. The true $n_i \times n_i$ covariance matrix of $\mathbf{y}_i$ given $\mathbf{X}_i$, $Var(\mathbf{y}_i \mid \mathbf{X}_i)$ is denoted by $\boldsymbol{\Omega}_i$. Let the maximum of n_i be q, and assume that q is bounded. The working correlation matrix for q repeated outcomes is denoted by $\mathbf{R}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is an $s \times 1$ vector fully characterizing $\mathbf{R}(\boldsymbol{\alpha})$. When the working correlation matrix is unstructured, $\boldsymbol{\alpha}$ can be $q^2 \times 1$ vectorized elements of $\mathbf{R}(\boldsymbol{\alpha})$. We denote by $\mathbf{R}_i(\boldsymbol{\alpha})$ the ith sub-matrix of $\mathbf{R}(\boldsymbol{\alpha})$ extracted according to the corresponding indices, and write $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{A}(\boldsymbol{\mu}_i)^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}(\boldsymbol{\mu}_i)^{1/2}$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})^T$, and $\mathbf{A}(\boldsymbol{\mu}_i)$ is a diagonal matrix with $a(\mu_{ij})$ as the jth diagonal element. Assume that we have $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$ that satisfy $K^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = O_p(1)$. The limit of $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}_0$, is determined by the value that satisfies the expectation of the estimating function for $\boldsymbol{\alpha}$ being zero. When the true and specified correlation structures are different, $\boldsymbol{\alpha}_0$ could be different depending on the estimating function for $\boldsymbol{\alpha}$, which leads to different asymptotic relative efficiency (Wang and Carey, 2003). A lack of definition of $\boldsymbol{\alpha}_0$ when working correlation is different from the true correlation is discussed in Crowder (1995). The GEE estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by solving GEE,

$$\mathbf{U}\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\} = \sum_{i=1}^{K} \mathbf{U}_i\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\} = \sum_{i=1}^{K} \mathbf{D}_i^T \boldsymbol{\Sigma}_i\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^T$. Let $\mathbf{W}_0(\boldsymbol{\beta}, \boldsymbol{\alpha}) = E(-K^{-1}\partial\mathbf{U}/\partial\boldsymbol{\beta}^T)$, $Var\{K^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\alpha})\}$ be $\mathbf{W}_1(\boldsymbol{\beta}, \boldsymbol{\alpha})$, where $\mathbf{W}_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}) = E\left\{K^{-1}\sum_{i=1}^{K} \mathbf{D}_i^T \boldsymbol{\Sigma}_i\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i\{\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\}^{-1}\mathbf{D}_i\right\}$. Notation $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ emphasizes that $\hat{\boldsymbol{\alpha}}$ is a function of $\boldsymbol{\beta}$. Under some conditions, $K^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is shown to be asymptotically normal with mean $\mathbf{0}$ and variance $\mathbf{W}_0^{-1}\mathbf{W}_1\mathbf{W}_0^{-1}$ (Liang and Zeger, 1986).

3. Motivation

To motivate the proposed method, we first quantify the loss of the asymptotic relative efficiency (ARE) by limiting the choice of working correlation structure to exchangeable and autoregressive of order 1 (AR-1). This quantification

Please cite this article in press as: Kwon, Y., et al., Generalized estimating equations with stabilized working correlation structure. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.08.016