Computational Statistics and Data Analysis 54 (2010) 1328–1341

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

ECM-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors

Wan-Lun Wang*, Tsai-Hung Fan

Graduate Institute of Statistics, National Central University, Jhongli 32001, Taiwan

ARTICLE INFO

Article history:
Received 7 July 2009
Received in revised form 9 November 2009
Accepted 24 November 2009
Available online 29 November 2009

ABSTRACT

For the analysis of longitudinal data with multiple characteristics, we are devoted to providing additional tools for multivariate linear mixed models in which the errors are assumed to be serially correlated according to an autoregressive process. We present a computationally flexible ECM procedure for obtaining the maximum likelihood estimates of model parameters. A score test statistic for testing the existence of autocorrelation among within-subject errors of each characteristic is derived. The techniques for the estimation of random effects and the prediction of further responses given past repeated measures are also investigated. The methodology is illustrated through an application to a set of AIDS data and two small simulation studies.

1. Introduction

During the past few decades, statistical methods for continuous longitudinal data with single-response repeated measures have received considerable attention via a vast amount of research. Laird and Ware (1982) developed the linear mixed model (LMM), which incorporates the random effects and allows for an unbalanced design in the sense that all subjects do not have an equal number of measurements and/or a common set of occasions. More general LMMs have been extensively studied by Jennrich and Schluchter (1986), Laird et al. (1987), Lindstrom and Bates (1988) and Chi and Reinsel (1989), among others. A comprehensive introduction to mixed models can be obtained from the monographs by Verbeke and Molenberghs (2000), Diggle et al. (2002), Fitzmaurice et al. (2004), and Hedeker and Gibbons (2006).

In many biomedical studies and clinical trials, it is quite common that repeated measures are collected on more than one response variable and are often referred to as multivariate longitudinal data. Reinsel (1984) proposed a general linear model with multivariate random effects to handle the balanced multi-outcome longitudinal data. Zucker et al. (1995) made inferences for the relationship between the subject-specific intercept and slope in a linear growth curve model. Shah et al. (1997) extended the LMM to the multivariate linear mixed model (MLMM), which allows analyzing unbalanced multivariate longitudinal data. Ideally, the MLMM has become the most important and frequently used analytical tool for continuous longitudinal data with multiple characteristics. Several alternative methods, including generalized estimating equations, two-stage factor analyses and the latent variable model, were considered in Sammel et al. (1999). Recently, Roy (2006) discussed how to estimate the correlation coefficient between two variables with repeated observations under bivariate linear mixed models.

Most works in the literature dealing with the MLMM have assumed that observations within each subject are serially uncorrelated. Since longitudinal data are occasionally collected over time, observations within each subject may tend to be serially correlated. To account for the effects of autocorrelation other than those caused by random effects, we exploit an autoregressive (AR) process of order p for the within-subject errors of each characteristic. Note that the use of a much

* Corresponding author. Tel.: +886 3 4227151; fax: +886 3 4258602.
E-mail address: wanlun@stat.ncu.edu.tw (W.-L. Wang).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.11.021

richer ARMA family is a straightforward extension (cf. Rochon, 1992 and Lee et al., 2005). In this paper, we aim at providing additional tools for the MLMM with AR(p) errors. The estimates of model parameters are calculated via the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), which is computationally flexible and conveniently implemented with existing softwares. Standard error estimates are calculated by inverting the Fisher information matrix.

In Section 2, we establish notation and present the model formulation. In Section 3, we discuss computational aspects of maximum likelihood (ML) estimation under a complete data framework. In Section 4, a score test statistic is offered to verify the existence of autocorrelation among the within-subject errors. Empirical Bayes estimation of random effects and prediction of future values are presented in Section 5. The proposed methodology is illustrated with a real example concerning the AIDS Clinical Trials Group Study 175 (ACTG 175) and two simulation studies in Section 6. Concluding remarks are given in Section 7, and the technical derivations are sketched in Appendix.

2. Model formulation

Suppose there are N subjects in a longitudinal study and for each subject there are r characteristics observed over time. Let $\mathbf{Y}_i = [\mathbf{y}_{i1} : \cdots : \mathbf{y}_{ir}]$ be an $n_i \times r$ matrix of response variables for subject i ($i = 1, \ldots, N$), where each $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_i})^{\mathrm{T}}$ is an $n_i \times 1$ vector of the jth ($j = 1, \ldots, r$) characteristic measured at particular time points $t = 1, \ldots, n_i$. For ease of notation, we use the vec() operator, which vectorizes a matrix by stacking its columns vertically; the vech() operator, which extracts the distinct elements of a matrix into a vector; and the Kronecker product, denoted by $\otimes$, which maps two arbitrarily dimensioned matrices into a larger matrix with a specific block structure.

We define a family of MLMM with AR(p) dependence as

$$\mathbf{Y}_i = \mathbf{X}_i\mathbf{A} + \mathbf{Z}_i\mathbf{B}_i + \mathbf{E}_i, \tag{1}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known full-rank covariate matrices of dimensions $n_i \times q_1$ and $n_i \times q_2$, respectively; $\mathbf{A} = [\boldsymbol{\alpha}_1 : \cdots : \boldsymbol{\alpha}_r]$ is a $q_1 \times r$ matrix of fixed effects with each vector of regression coefficients $\boldsymbol{\alpha}_j$ being used to describe the population mean of each corresponding characteristic; $\mathbf{B}_i$ is a $q_2 \times r$ matrix of unobservable random effects; and $\mathbf{E}_i = [\mathbf{e}_{i1} : \cdots : \mathbf{e}_{ir}]$ is an $n_i \times r$ matrix of residuals. For notation simplicity, we shall replace vec($\mathbf{B}_i$), vec($\mathbf{E}_i$) and vec($\mathbf{A}$) with $\mathbf{b}_i$, $\mathbf{e}_i$ and $\boldsymbol{\alpha}$, respectively. Furthermore, we assume $\mathbf{b}_i \sim N_{q_2 r}(\mathbf{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_{jj'}]$ is a $q_2 r \times q_2 r$ positive-definite matrix with $\boldsymbol{\psi}_{jj'}$'s ($j, j' = 1, \ldots, r$) being $q_2 \times q_2$ block partitioned matrices. In particular for $j = j'$, $\boldsymbol{\psi}_{jj}$ can be viewed as a covariance structure of the random effects for the jth characteristic only. For the within-subject errors, we assume $\mathbf{e}_i \sim N_{n_i r}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{C}_i)$, independent of $\mathbf{b}_i$'s, where $\boldsymbol{\Sigma} = [\sigma_{jj'}]$ is an $r \times r$ unstructured matrix describing the variance and covariance among r response variables, and $\mathbf{C}_i$ is an $n_i \times n_i$ structured AR(p)-process matrix. The form of $\mathbf{C}_i = \mathbf{C}_i(\boldsymbol{\phi}) = [\rho_{|t-t'|}(\boldsymbol{\phi})]$, for $t, t' = 1, \ldots, n_i$, is specified to address the autocorrelation among n_i occasions on each outcome, where

$$\rho_s(\boldsymbol{\phi}) = \rho_s = \phi_1\rho_{s-1} + \cdots + \phi_p\rho_{s-p}, \quad \rho_0 = 1, \ (s = 0, \ldots, n_i - 1),$$

namely the Yule–Walker equation (Box et al., 1994), is an implicit function of the AR parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^{\mathrm{T}}$. For the pure AR model, the admissible values of $\boldsymbol{\phi}$ are restricted in a p-dimensional hypercube $\mathbb{C}^p$. To ensure the stationarity of the AR model, the roots of $1 - \phi_1\mathbb{B} - \phi_2\mathbb{B}^2 - \cdots - \phi_p\mathbb{B}^p = 0$ must lie outside the unit circle, where $\mathbb{B}$ is a backward shift operator such that $\mathbb{B}^v\rho_s = \rho_{s-v}$, for $v = 0, \ldots, p$.

From model (1), the joint distribution of vec($\mathbf{Y}_i$) and $\mathbf{b}_i$ is

$$\begin{bmatrix} \mathrm{vec}(\mathbf{Y}_i) \\ \mathbf{b}_i \end{bmatrix} \overset{\text{ind}}{\sim} N_{(n_i+q_2)r}\left(\begin{bmatrix} \mathrm{vec}(\mathbf{X}_i\mathbf{A}) \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Lambda}_i & (\mathbf{I}_r \otimes \mathbf{Z}_i)\boldsymbol{\Psi} \\ \boldsymbol{\Psi}(\mathbf{I}_r \otimes \mathbf{Z}_i)^{\mathrm{T}} & \boldsymbol{\Psi} \end{bmatrix} \right), \tag{2}$$

where $\boldsymbol{\Lambda}_i = (\mathbf{I}_r \otimes \mathbf{Z}_i)\boldsymbol{\Psi}(\mathbf{I}_r \otimes \mathbf{Z}_i)^{\mathrm{T}} + \boldsymbol{\Sigma} \otimes \mathbf{C}_i$. We remark that the LMM with AR(1) of Chi and Reinsel (1989), specified by $\mathbf{y}_{ij} \sim N_{n_i}(\mathbf{X}_i\boldsymbol{\alpha}_j, \mathbf{Z}_i\boldsymbol{\psi}_{jj}\mathbf{Z}_i^{\mathrm{T}} + \sigma_{jj}\mathbf{C}_i(\phi_1))$ with $\mathbf{C}_i(\phi_1)$ taking the form of $(1 - \phi_1^2)^{-1}[\phi_1^{|t-t'|}]$ ($t, t' = 1, \ldots, n_i$; $-1 < \phi_1 < 1$), can be treated as a special case of model (1). Let $\boldsymbol{\theta} = (\mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$ denote the entire model parameters. It follows from (2) that the log-likelihood function of $\boldsymbol{\theta}$ for $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, omitting the constant term, is

$$\ell = \ell(\boldsymbol{\theta}|\mathbf{Y}) = -\frac{1}{2}\sum_{i=1}^{N}\Big\{\log|\boldsymbol{\Lambda}_i| + \mathrm{vec}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{A})^{\mathrm{T}}\boldsymbol{\Lambda}_i^{-1}\mathrm{vec}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{A})\Big\}. \tag{3}$$

Explicit expressions for the score vector $\mathbf{s}_{\boldsymbol{\theta}}$ and the Fisher information matrix $\mathbf{J}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ for $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\omega} = (\mathrm{vech}(\boldsymbol{\Psi})^{\mathrm{T}}, \mathrm{vech}(\boldsymbol{\Sigma})^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$, are derived in Appendix A.

3. Maximum likelihood estimation via the ECM algorithm

3.1. Parameter estimation

The EM algorithm (Dempster et al., 1977) has been well recognized as a useful tool for ML estimation in models with missing values or latent data. The ECM algorithm is a generalization of EM in which the maximization (M) step is replaced by