Journal of Multivariate Analysis 143 (2016) 481–491

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

The analysis of multivariate longitudinal data using multivariate marginal models

Hyunkeun Cho

Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, United States

ARTICLE INFO

Article history:
Received 9 November 2014
Available online 1 November 2015

AMS subject classifications:
62H12
62H15

Keywords:
Generalized estimating equation
Longitudinal data
Multiple responses
Multivariate marginal models
Quadratic inference function

ABSTRACT

Longitudinal studies often involve multiple outcomes measured repeatedly from the same subject. The analysis of multivariate longitudinal data can be challenging due to its complex correlated nature. In this paper, we develop multivariate marginal models in longitudinal studies with multiple response variables, and improve parameter estimation by incorporating informative correlation structures. In theory, we show that the proposed method yields a consistent and efficient estimator which follows an asymptotic normal distribution. Monte Carlo studies indicate that the proposed method performs well in the sense of reducing bias and improving estimation efficiency. In addition, the proposed approach is applied to a real longitudinal data example of transportation safety with different response families.

1. Introduction

Multivariate longitudinal analysis has increased in popularity in several disciplines where subjects are measured across time with regard to a collection of response variables. Multivariate longitudinal data provides a unique opportunity in studying the joint evolution of various responses over a period of time. Unlike the traditional longitudinal studies with the univariate response, the analysis of multivariate longitudinal data can be challenging because both repeated measurements from the same subject and different response variables are likely to be correlated.

The correlated nature of longitudinal data often makes it difficult to specify the full likelihood function. The generalized estimating equation (GEE, [12]) is a suitable approach for parameter estimation for longitudinal data without specification of the likelihood. Although the GEE yields a consistent estimator and variance estimates, the estimator can be inefficient under the incorrect specification of the correlation structure. Qu, Lindsay and Li [13] developed the quadratic inference function (QIF) to improve the efficiency of the GEE when the working correlation is incorrectly specified. However, the QIF approach ignores the multivariate response association and constructs univariate marginal models for parameter estimation. Model accuracy and efficiency can be improved by incorporating the correlation information among responses effectively.

Repeated observations of multivariate response variables require a multivariate longitudinal framework. Generalized linear mixed models have been extended in multivariate longitudinal studies; see [9,4,10,1,14]. When the number of parameters increases with the sample size and a collection of responses, the random-effect approaches are more likely to be computationally intensive and unstable. In addition, it is difficult to evaluate the marginal likelihood of jointly generalized linear mixed models when the response is non-normal.

Contrary to mixed-models approaches, Asar and İlk [3] utilize the generalized estimating equation based on multiple marginal models of multiple responses. However, the asymptotic properties such as consistency and efficiency of estimation

E-mail address: hyunkeun.cho@wmich.edu.

http://dx.doi.org/10.1016/j.jmva.2015.10.012
0047-259X/

have not been studied. In practice, the construction of the joint marginal model becomes more complex when we impose on certain aspects of the correlation structure on multivariate longitudinal analysis. In addition, inferences of interest are easily influenced by the correlation structure's assumptions. Alternatively, if the unstructured correlation structure is considered, it might cause convergence problems as the number of parameters to be estimated grows rapidly [3]. Furthermore, the marginal modeling approach is mainly applicable for a collection of the same response family.

In this paper, we provide the estimation procedure for multiple longitudinal data by using the quadratic inference function approach for multivariate marginal models. The proposed approach is able to estimate all parameters corresponding to multiple responses simultaneously even if the type of the responses are different. In addition, the proposed method can easily take correlation information from repeated measures within the subject and among different responses without estimating the correlation parameters, yet it does not require specifying the likelihood functions. Our theoretical investigations and simulation studies show that the estimator of the proposed approach is consistent and more efficient than the estimator obtained by a certain amount of correlation information. Furthermore, the proposed approach also provides an inference function for model diagnostic tests and goodness-of-fit tests for multivariate longitudinal data. The multivariate modeling approach is applied on a real longitudinal data set on the transportation safety study that consists of a discrete response variable (the crash frequency) and a binary response variable (the presence of crash severity).

The paper is organized as follows. In Section 2, we propose the estimation procedure and statistical inferences for multivariate longitudinal responses, illustrate the choice of the correlation structure, and discuss how to implement for data with missing. Section 3 provides simulation studies and data analysis for the transportation safety study. We conclude remarks with a brief discussion in Section 4. The theoretical proofs are placed in the Appendix.

2. Longitudinal data analysis with multivariate responses

2.1. Estimation procedure for multivariate marginal models

Suppose $\mathbf{y}_{i\cdot k} = (y_{i1k}, \ldots, y_{ij_ik})'$ is the kth response variable measured J_i times from the ith subject, and y_{ijk}'s are independent identically distributed for $i = 1, \ldots, N$, where N is the sample size and J_i is the cluster size. To simplify the notation, we first set $J_i = J$ for all i and the unbalanced data case will be discussed with more details in Section 2.4. For the generalized linear model, the formulation of multivariate marginal models is defined as

$$\mu_{ijk} = E(y_{ijk}|\mathbf{x}_{ij}) = \mu(\mathbf{x}_{ij}'\boldsymbol{\beta}_k), \tag{1}$$

where $\mu(\cdot)$ is an inverse link function, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kP})'$ is a P-dimensional parameter vector for the kth response and $\mathbf{x}_{ij}$ is the corresponding covariate at time j for the ith subject.

To accommodate the association between responses, we stack up the response variable as $\mathbf{Y}_i = (\mathbf{y}_{i\cdot 1}', \ldots, \mathbf{y}_{i\cdot K}')'$ and $\mathbf{X}_i = (\mathbf{I}_K \otimes \mathbf{x}_i)$ is extended to a $PK \times JK$ matrix by Kronecker product operator, where K is the number of responses, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})$ is a $P \times J$ matrix, $\mathbf{I}_K$ is a $K \times K$ identity matrix and $\otimes$ corresponds to a left Kronecker product. The corresponding parameter is a PK-dimensional vector of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K')'$ and the marginal model in (1) is represented as $\boldsymbol{\mu}_i = E(\mathbf{Y}_i|\mathbf{X}_i) = \mu(\mathbf{X}_i'\boldsymbol{\beta})$. We extend the quasi-likelihood to incorporate the correlation information and obtain the estimator by solving

$$\sum_{i=1}^{N} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2} \mathbf{R}(\boldsymbol{\alpha})^{-1} \mathbf{A}_i^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{2}$$

where $\dot{\boldsymbol{\mu}}_i = (\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta})$, $\mathbf{A}_i$ is the $JK \times JK$ diagonal marginal variance matrix of $\mathbf{Y}_i$, and $\mathbf{R}(\boldsymbol{\alpha})$ is the working correlation matrix that contains correlation parameters $\boldsymbol{\alpha}$. The approach requires only a few nuisance parameters $\boldsymbol{\alpha}$ to specify a common working correlation structure such as an exchangeable or the first-order autoregressive (AR1) correlation.

For the multivariate marginal model, the working correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ enables us to accommodate three pieces of association information; the correlation across time within the subject, the cross-correlation between different response variables both at the same time and across time. Therefore, the simple working correlation structure such as exchangeable or AR1 does not represent the true correlation structure sufficiently. It is well-known that when the correlation structure is incorrectly specified, the estimator can be inefficient. If the unspecified correlation structure is considered as the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$, there are $(JK) \times (JK-1)/2$ correlation parameters $\boldsymbol{\alpha}$ to be estimated, which might cause convergence problems when the cluster size is large.

To avoid the estimation of $\boldsymbol{\alpha}$, Qu, Lindsay and Li [13] formulate the inverse of $\mathbf{R}$ by a linear combination of basis matrices,

$$\mathbf{R}^{-1} = b_0 \mathbf{I} + \sum_{m=1}^{q} b_m \mathbf{B}_m, \tag{3}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{B}_1, \ldots, \mathbf{B}_q$ are basis matrices with 0 and 1 components and b_m's are unknown coefficients. The choice of basis matrices will be discussed in more detail in Section 3. By replacing $\mathbf{R}(\boldsymbol{\alpha})^{-1}$ in (2) with basis matrices in (3),