Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Simultaneous variable selection and estimation for multivariate multilevel longitudinal data with both continuous and binary responses[☆]

Haocheng Li [a,b,*], Di Shu [c], Yukun Zhang [a], Grace Y. Yi [c]

[a] *Department of Oncology, University of Calgary, Canada*
[b] *Department of Community Health Sciences, University of Calgary, Canada*
[c] *Department of Statistics and Actuarial Science, University of Waterloo, Canada*

ARTICLE INFO

Article history:
Received 2 January 2017
Received in revised form 6 September 2017
Accepted 6 September 2017
Available online xxxx

Keywords:
Longitudinal data
Mixed effects model
Multivariate multilevel longitudinal data
Penalized quasi-likelihood
Variable selection

ABSTRACT

Complex structured data settings are studied where outcomes are multivariate and multilevel and are collected longitudinally. Multivariate outcomes include both continuous and discrete responses. In addition, the data contain a large number of covariates but only some of them are important in explaining the dynamic features of the responses. To delineate the complex associate structures of the responses, a model with correlated random effects is proposed. To handle the large dimensionality of covariates, a simultaneous variable selection and parameter estimation method is developed. To implement the method, a computationally feasible algorithm is described. The proposed method is evaluated empirically by simulation studies and illustrated by analyzing the data arising from the Waterloo Smoking Prevention Project.

1. Introduction

The Waterloo Smoking Prevention Project (WSPP) is a longitudinal study that collects smoking status, psychological measures and social background information by following Canadian middle school students from Grades 6 to 12 through an annual questionnaire survey (Cameron et al., 1999). The survey is implemented by first selecting a list of schools and then taking a group of students within each school to answer the questionnaire. A major research interest is the influence of smoking intervention, interpersonal relationship and social background factors on six outcomes: level of rebellion activity, level of weariness of study, level of self-confidence, current smoking status, smoking preference in the future, and intention to become thinner, in which three are continuous outcomes and three are binary outcomes. Researchers believe these outcomes are correlated, and it is interesting to study how these outcomes are correlated. Furthermore, it is perceived that not all collected covariates are useful in explaining the changes in the outcomes, and thus, it is imperative to exclude those irrelevant covariates when building models and performing inferences.

Motivated by the features of the WSPP data, we propose a new methodology to handle complex longitudinal data that consist of (1) multiple responses including both continuous and binary outcomes, (2) multilevel structures with clusters, and (3) a large number of covariates in which not all of them are useful for explaining responses. Each of these features has individually received much attention in the literature, but, as far as we know, methodology does not exist

[☆] Supplementary material is available online.
[*] Correspondence to: Departments of Oncology and Community Health Sciences, University of Calgary, Alberta, Canada T2N 4N2.
E-mail address: haocheng.li@ucalgary.ca (H. Li).

http://dx.doi.org/10.1016/j.csda.2017.09.004
0167-9473/© 2017 Elsevier B.V. All rights reserved.

to handle these characteristics simultaneously. Gueorguieva and Agresti (2001) discussed joint models to address the problem with one continuous response and one binary outcome, but their method cannot be directly applied for multilevel data. Gibbons and Hedeker (1997) and Li et al. (2015) proposed multilevel models to deal with either a binary response or a continuous response, but not both. Goldstein (2011) considered a class of models for multivariate multilevel data, but their estimation algorithm is computationally expensive when the number of random effects is large. In terms of variable selection with both fixed and random effects, Bondell et al. (2010) and Ibrahim et al. (2010) explored selection methods via the EM algorithm. Groll and Tutz (2014), Pan and Huang (2014), and Hui et al. (2017a, in press) proposed variable selection approaches for generalized linear mixed models. However, these methods cannot handle the data with multivariate multilevel structures.

Our goal is to develop a flexible statistical method that handles complex structured longitudinal data. We propose a multivariate multilevel mixed effects model for continuous and binary outcomes, where the correlation across responses and levels are postulated by random effects. Inferences under our proposed model, however, cannot be carried out using the traditional likelihood method (e.g. Longford, 1994), because the number of random effects would rapidly increase as the number of outcomes and the number of subjects within clusters become large. Fieuws and Verbeke (2006) pointed out that even in the case of multivariate normal data where the likelihood function assumes a manageable form, computation is difficult if the number of outcomes exceeds four due to the rapid increment of parameters in the covariance matrix of random effects. For models with multiple binary responses, integration problems in likelihood function are even more difficult to handle when the dimension of random effects is large (Fieuws et al., 2006). Although Gibbons and Hedeker (1997) commented that the dimension of the integrals can be reduced under certain assumptions, the computation complexity and intensity are still a challenge for general settings with more than three random effects. In our proposed model, large dimensional random effects are required to feature multivariate and multilevel longitudinal data, which present a great challenge in computation.

A naive solution is to ignore the correlation among all the outcomes and treat them independently, but this is apparently infeasible in dealing with various practical problems. Modeling the association patterns among all continuous and binary responses is important in many applications, and ignoring such correlation structures would degrade inference results. An alternative approach is to use a pairwise strategy by respectively fitting paired combinations among all the outcomes (Fieuws and Verbeke, 2006; Li and Yi, 2013). However, this method can still be infeasible as the number of pairs increases quadratically with the number of outcomes. In addition, if a pair involves two binary responses, the dimension for random effects may still be difficult for numerical integrations.

Groll and Tutz (2014), Pan and Huang (2014) and Hui et al. (2017a, in press) examined a penalized quasi-likelihood method to handle generalized linear mixed models. However, their approach often leads to biased estimates as pointed out by Rodriguez and Goldman (2001). To get around the aforementioned issues, we propose an estimation method by employing the augmented penalized quasi-likelihood framework (Breslow and Clayton, 1993) and utilizing a second-order Taylor series approximation (Goldstein and Rasbash, 1996). Our approach extends existing models by accommodating complex data structures with more flexible random effect structures introduced. We apply transformations to binary responses so that the transformed outcomes can be approximately postulated by linear mixed models. Such a treatment allows all the outcomes to be jointly estimated by an approximated multivariate multilevel linear mixed model, which in turn, leads to a manageable computation algorithm. The idea of the Expectation/Conditional Maximization Either (ECME) algorithm (Schafer, 1998) is extended to our settings to efficiently handle the estimation of the parameters induced from the large dimension of random effects. Our proposed method overcomes the computation inability of the usual likelihood method for handling complex structured data and the estimation limitations of the pairwise modeling strategies.

In practice, irrelevantly incorporating a large number of fixed and random effects into the model may result in the difficulty of computation, interpretation and prediction, thus parsimonious models are typically desirable. We further augment the proposed algorithm to handle the estimation of model parameters as well as selecting appropriate fixed and random effects.

The paper is organized as follows. Section 2 describes the hierarchical model for multiple continuous and binary variables. In Section 3, we develop a feasible computation algorithm for estimation. In Section 4, we propose the variable selection approach. In Section 5, we assess the performance of the proposed method via simulation studies, and in Section 6 we analyze the motivating data using the proposed method. Concluding remarks are presented in Section 7.

2. Notation and framework

For $\ell = 1, \ldots, M$, let $Y_{ijk}^{(\ell)}$ be the ℓth outcome measured at occasion $k = 1, \ldots, K_{ij}$ for subject $j = 1, \ldots, J_i$ in cluster $i = 1, \ldots, n$. For ease of exposition, let $\mathbf{C}$ and $\mathbf{B}$ denote the index set for continuous responses and binary responses, respectively.

Mixed effect models for continuous and binary outcomes are specified as

$$Y_{ijk}^{(\ell)} = \eta_{ijk}^{(\ell)} + \epsilon_{ijk}^{(\ell)}, \tag{1}$$

and

$$\text{logit}[\text{pr}\{Y_{ijk}^{(\ell)} = 1 | u_i^{(\ell)}, \nu_{ij}^{(\ell)}\}] = \eta_{ijk}^{(\ell)}, \tag{2}$$

Please cite this article in press as: Li H., et al., Simultaneous variable selection and estimation for multivariate multilevel longitudinal data with both continuous and binary responses. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.09.004.