Statistica Sinica **21** (2011), 207-229

A PAIRWISE LIKELIHOOD METHOD FOR CORRELATED BINARY DATA WITH/WITHOUT MISSING OBSERVATIONS UNDER GENERALIZED PARTIALLY LINEAR SINGLE-INDEX MODELS

Wenqing He and Grace Y. Yi

University of Western Ontario and University of Waterloo

Abstract: Correlated data, such as multivariate or clustered data, arise commonly in practice. Unlike analysis for independent data, valid inference based on such data often requires proper accommodation of complex association structures among response components within clusters. Semiparametric models based on generalized estimating equations (GEE) methods, and their extensions, have become increasingly popular. However, these inferential schemes are greatly challenged by the complexity of such data features as missing observations, ubiquitous in applications. Moreover, existing methods mainly concern marginal mean parameters with association parameters treated as nuisance. This treatment is inadequate to handle clustered data for which estimation of association parameters can be a central theme of the study. To address these problems, we develop a flexible semiparametric method that can handle correlated data with or without missing values. Our discussion focuses on binary data that arise commonly. The proposed method enjoys a number of attractive properties, including that the missing data process is left unmodeled, yet model assumptions for the response process are kept to a minimum. It is robust in the sense that only the mean and association structures for the response process are modeled. The proposed method is flexible because both parametric and nonparametric structures are incorporated in modeling the mean responses.

Key words and phrases: Association, binary outcomes, correlated data, generalized partially linear models, missing data, pairwise likelihood, semiparametric method, single-index models.

1. Introduction

Correlated data, including clustered data, multivariate data, and longitudinal data, arise commonly in applications. A number of inference methods have been developed for handling such data. In particular, the generalized estimating equations (GEE) approach is widely used for analysis of longitudinal or clustered data (Liang and Zeger (1986)). This marginal approach is viewed as attractive because it does not require complete specification of the joint distribution of the correlated responses, but rather is based only on specification of their first two moments. While primary interest most frequently lies in making inference about the parameters in regression models for the marginal means, there has been increasing interest in estimation of association parameters (e.g., Connolly and Liang (1988)). When the association parameters are of central importance, second order generalized estimating equations can be constructed. With binary data Prentice (1988) developed such equations and emphasized estimation of the correlation parameters. Fitzmaurice and Laird (1993) proposed a model which parameterizes the association in terms of conditional odds ratios, whereas Lipsitz, Laird, and Harrington (1991), Liang, Zeger, and Qaqish (1992) Carey, Zeger and Diggle (1993), Molenberghs and Lesaffre (1994), Lang and Agresti (1994), and Fitzmaurice and Lipsitz (1995) proposed models which parameterize the association in terms of marginal odds ratios.

Those methods apply to data that contain no missing values. However, it is often the case that missing observations are present with correlated data. When data are missing completely at random (MCAR), the GEE approach based on the observed data can still produce consistent estimators for the response parameters because missing data processes are not related to the processes generating responses. In contrast, when data are missing at random (MAR) or missing not at random (MNAR), the estimating equations are not unbiased and hence fail to provide consistent estimates. The inverse probability weighted GEE (IPWGEE) approach has been developed (Robins, Rotnitzky, and Zhao (1995)) to conduct valid inference under MCAR or MAR mechanisms. The IPWGEE method and its extensions have been discussed extensively in the literature, see Robins, Rotnitzky, and Zhao (1995), Fitzmaurice, Molenberghs, and Lipsitz (1995), and Yi and Cook (2002a,b), for example.

The validity of the IPWGEE approach relies on correctly modeling the missing data process. In practice, however, it is generally difficult to tell from data what missing data mechanism is reasonable and what model might correctly characterize the missing data process. If the missing data process is incorrectly postulated, then the resulting inference on the response parameters may be seriously biased. To overcome this problem, we may alternatively adopt the likelihood approach which leads to, under MAR mechanisms, valid inference by simply using the observed data. The advantage of this approach is that the missing data process is left unattended. However, this advantage is achieved at the price of fully specifying the joint distribution of the response process which, in many cases, and especially for multivariate discrete responses, is not a trivial task.

It is desirable to have a method that combines the advantages of both likelihood and GEE approaches. To this end, we explore using the pairwise likelihood approach (Lindsay (1988); Cox and Reid (2004)) to handle correlated binary data that may be complete or incomplete. The proposed method preserves appealing features of both likelihood and GEE approaches. Thus, our method requires only minimal model assumptions for the response process like the GEE approach, yet it allows the missing data process to be left unspecified just as the likelihood method does. In addition, the proposed method facilitates estimation of both mean and association parameters.

To accommodate a richer class of mean structures, we allow flexibility in modeling of the response process by using a semiparametric regression to postulate mean structures. Specifically, we employ generalized partially linear single-index models to feature mean responses (e.g., Carroll et al. (1997); Yi, He, and Liang (2009)). Such models are useful when the commonly adopted linear relationship between the mean response and covariates, under a suitable link, becomes inadequate. However, such modeling flexibility induces considerable challenges in estimation procedures and the establishment of asymptotic properties. The computing algorithm for usual estimating equations based on the Newton-Raphson method cannot be employed directly due to the inclusion of nonlinear functions whose forms are not known. To circumvent this problem, we use the local polynomial smoothing technique (Fan, Heckman, and Wand (1995)) to perform estimation.

The remainder of the paper is organized as follows. In Section 2 we introduce notation and model assumptions for the subsequent discussion. In Section 3 we present the estimation and inference procedures for analyzing correlated data with or without missing values. Empirical assessment of the proposed method is reported in Section 4, along with an application. General remarks are made in Section 5.

2. Notation and Framework

Suppose there are n clusters and m_i subjects within cluster i, $i = 1, \ldots, n$. Let Y_{ij} be the binary response for subject j in cluster i, which may be missing; $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ be the covariate vectors. Let $R_{ij} = I(Y_{ij} \text{ is observed})$ be the missing data indicator, where $I(.)$ is the indicator function. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})^T$, $\mathbf{z}_i = (\mathbf{z}_{i1}, \ldots, \mathbf{z}_{im_i})^T$, and $\mathbf{R}_i = (R_{i1}, \ldots, R_{im_i})^T$. Take $\mu_{ij} = E(Y_{ij}|\mathbf{x}_i, \mathbf{z}_i)$, and let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im_i})^T$, $i = 1, \ldots, n$. We consider the regression model

$$g^{-1}(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \theta(\mathbf{z}_{ij}^T \boldsymbol{\alpha}) \quad \text{with } \|\boldsymbol{\alpha}\| = 1, \tag{2.1}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unknown parameter vectors, $\theta(\cdot)$ is an unknown smooth function, and $g(\cdot)$ is a known monotone link function. Common choices of g include logit, probit, or complement log-log functions. The requirement $\|\boldsymbol{\alpha}\| = 1$ ensures the identifiability of $\boldsymbol{\alpha}$ (Carroll et al. (1997)).

We assume that Y_{ij} and $Y_{i'j'}$ are independent for different clusters i and i', but within the same cluster, may be correlated. Various measures have been proposed to quantify the association between binary outcomes. For example, Prentice (1988) discussed using correlation coefficients for measuring association for longitudinal binary data, and Zhao and Prentice (1990) discussed a measure based on covariances. Odds ratios, on the other hand, have received increasing interest partly due to the fact that there is no constraint associated with such measures. Specifically, conditional odds ratios (e.g., Fitzmaurice and Laird (1993)) and marginal odds ratios (e.g., Lipsitz, Laird, and Harrington (1991)) have been widely used. As conditional odds ratios may not have a convenient interpretation independent of the cluster size, we focus our discussion on marginal odds ratios. Let ψ_{ijk} be the marginal odds ratio between responses Y_{ij} and Y_{ik} in the same cluster i,

$$\psi_{ijk} = \frac{P(Y_{ij}=1, Y_{ik}=1|\mathbf{x}_i, \mathbf{z}_i) \cdot P(Y_{ij}=0, Y_{ik}=0|\mathbf{x}_i, \mathbf{z}_i)}{P(Y_{ij}=1, Y_{ik}=0|\mathbf{x}_i, \mathbf{z}_i) \cdot P(Y_{ij}=0, Y_{ik}=1|\mathbf{x}_i, \mathbf{z}_i)}, \; j \neq k.$$

Regression models may be employed to characterize various association structures, with the dependence of the association on the covariates being explicitly reflected. Typically, a log linear regression may be assumed with

$$\log \psi_{ijk} = \mathbf{u}_{ijk}^T \boldsymbol{\phi}, \tag{2.2}$$

where $\mathbf{u}_{ijk}$ is a vector of covariates that specifies the form of the association between Y_{ij} and Y_{ik}, and $\boldsymbol{\phi}$ is a vector of regression parameters. Letting $\mathbf{u}_{ijk}$ be the scalar one, for example, leads to the exchangeable association between responses within the same cluster, while setting $\mathbf{u}_{ijk}^T \boldsymbol{\phi} = \phi^{|j-k|}$ results in autogressive correlation among responses.

Let $\mu_{ijk} = P(Y_{ij}=1, Y_{ik}=1|\mathbf{x}_i, \mathbf{z}_i)$ be the joint probability for the pair (Y_{ij}, Y_{ik}), given the covariates $\mathbf{x}_i$ and $\mathbf{z}_i$. This is determined by the marginal means and the odds ratio, given by (e.g., Lipsitz, Laird, and Harrington (1991)); Yi and Thompson (2005)).

$$\mu_{ijk} = \begin{cases} \dfrac{a_{ijk} - \sqrt{b_{ijk}}}{2(\psi_{ijk}-1)}, & \text{if } \psi_{ijk} \neq 1, \\ \mu_{ij}\mu_{ik}, & \text{if } \psi_{ijk} = 1, \end{cases}$$

where $a_{ijk} = 1 - (1-\psi_{ijk})(\mu_{ij}+\mu_{ik})$ and $b_{ijk} = a_{ijk}^2 - 4\psi_{ijk}(\psi_{ijk}-1)\mu_{ij}\mu_{ik}$.

3. Inference Procedures

3.1. Estimation algorithm

To estimate mean parameters $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ and the association parameter $\boldsymbol{\phi}$, one can employ the generalized estimating equations (GEE) approach as in, for example, Prentice (1988) and Yi and Cook (2002a,b) when the mean structure is modeled parametrically, and in Yi, He, and Liang (2009, 2010) when the mean structure is specified semi-parametrically. Typically, Prentice (1988) and Yi, He, and Liang (2009) deal with complete data; while Yi and Cook (2002a,b) and Yi, He, and Liang (2010) consider data with missing observations where a particular model for the missing data process is required.

Alternatively, we propose a likelihood-related approach that can handle both complete and incomplete data in a unified framework. Model assumptions for the response process are the same as those required by the GEE approach, but there is no need to specify a model for the missing data process. For $j \neq k$, let $L_{ijk} = P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}|\mathbf{x}_i, \mathbf{z}_i)$ be the joint probability for paired responses (Y_{ij}, Y_{ik}); this is determined by the marginal probability (2.1) and the odds ratio (2.2). Let $\mathcal{B} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\phi}^T)^T$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{im_i})^T$, with $\eta_{ij} = \theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})$, $j = 1, \ldots, m_i$. Let $\mathcal{L}_i(\boldsymbol{\eta}_i, \mathcal{B}) = \log\{\prod_{j<k} L_{ijk}^{R_{ij}R_{ik}}\}$, $\mathcal{L}_{ij}(\boldsymbol{\eta}_i, \mathcal{B}) = \partial\mathcal{L}_i(\boldsymbol{\eta}_i, \mathcal{B})/\partial\eta_{ij}$, and $\mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_i, \mathcal{B}) = \partial\mathcal{L}_i(\boldsymbol{\eta}_i, \mathcal{B})/\partial\mathcal{B}$, $j = 1, \ldots, m_i$. To incorporate varying numbers of observed components in different clusters, let $O_i = \sum_{j=1}^{m_i} I(R_{ij} = 1)$ denote the number of the observed measurements in cluster i, and $\mathbf{W}_i = w_i\mathbf{I}_{r\times r}$, where $w_i = 1/(O_i - 1)$, and $\mathbf{I}_r$ represents a $r \times r$ identity matrix with r denoting the dimension of parameter $\mathcal{B}$. Set $\mathbf{S}_{i\mathcal{B}} = \mathbf{W}_i \cdot \mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_i, \mathcal{B})$ and $S_{ij} = w_i \cdot \mathcal{L}_{ij}(\boldsymbol{\eta}_i, \mathcal{B})$. In Appendix 1 we show the following.

Theorem 1.

(a) *If the distributions* $f(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)$ *and* $f(\mathbf{r}_i|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i)$ *for the missing data indicator vector* $\mathbf{R}_i$ *do not depend on the response parameters* $\mathcal{B}$ *for any* $j, k = 1, \ldots, m_i$*, then* $\mathbf{S}_{i\mathcal{B}}$ *has zero mean.*

(b) *If the distributions* $f(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)$ *and* $f(\mathbf{r}_i|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i)$ *for the missing data indicator vector* $\mathbf{R}_i$ *do not depend on* $\boldsymbol{\eta}_i$ *for any* $j, k = 1, \ldots, m_i$*, then* S_{ij} *has zero mean,* $j = 1, \ldots, m_i$*.*

If $\theta(\cdot)$ is a known function, consistent estimators of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\phi}$ can be obtained, as discussed in Yi, Zeng and Cook (2010), by solving $\sum_{i=1}^n \mathbf{S}_{i\mathcal{B}} = \mathbf{0}$ for $\mathcal{B}$, due to Theorem 1 (a). Here, however, $\theta(\cdot)$ is unknown, hence we need to use nonparametric approaches to estimate it locally in order to estimate $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\phi}$. Assuming $\theta(u)$ has a second derivative for any given u, we approximate $\theta(u)$ by a linear function within the neighborhood of u_0 via the Taylor series expansion

$$\theta(u) \approx \theta(u_0) + \theta^{(1)}(u_0)(u - u_0)$$

for a point u_0 in the interior of the support of $\theta(\cdot)$, where $d^{(j)}(\cdot)$ denotes the jth derivative for function $d(\cdot)$. Let $K(u)$ be a kernel function (or a symmetric density function) with a compact support, and h be a bandwidth. Write $K_h(t) = K(t/h)/h$. Let $a_0(u_0) = \theta(u_0), a_1(u_0) = h\theta^{(1)}(u_0)$, $\mathbf{a}(u_0) = (a_0(u_0), a_1(u_0))^T$, $U_{ij} = \mathbf{z}_{ij}^T\boldsymbol{\alpha}$, and $\mathbf{G}_{ij}(u) = \{1, (U_{ij} - u)/h\}^T$.

To estimate $\mathcal{B}$, we propose the profile kernel method. That is, we first estimate the values of $\theta(\cdot)$ and $\theta^{(1)}(\cdot)$ over a selected grid for any given $\mathcal{B}$ using the kernel method, then we obtain the profile estimator $\widehat{\mathcal{B}}_p$ with $\theta(\cdot)$ and its derivative $\theta^{(1)}(\cdot)$ being fixed at their kernel estimates. To be specific, we proceed with the following two-stage estimation procedure.

- **Stage 1**. Let $\widehat{\boldsymbol{\eta}}_{i(j)}(\mathcal{B},u) = (\widehat{\theta}(U_{i1},\mathcal{B}),\ldots,\widehat{\theta}(U_{i,j-1},\mathcal{B}), a_0(u) + a_1(u)(U_{ij} - u)/h, \widehat{\theta}(U_{i,j+1},\mathcal{B}), \ldots, \widehat{\theta}(U_{im_i},\mathcal{B}))^T, j = 1,\ldots,m_i$. Given $\mathcal{B}$, for a given point u in the selected grid find $\widehat{\theta}(u,\mathcal{B}) = \widehat{a}_0(u)$ and $\widehat{\theta}^{(1)}(u,\mathcal{B}) = \widehat{a}_1(u)/h$ by solving

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(U_{ij}-u)\mathbf{G}_{ij}(u,h)\cdot w_i\mathcal{L}_{ij}(\boldsymbol{\eta}_i,\mathcal{B})|_{\boldsymbol{\eta}_i=\widehat{\boldsymbol{\eta}}_{i(j)}(\mathcal{B},u)} \tag{3.1}$$

  with respect to $\mathbf{a}(u)$.
  To solve (3.1), we may follow a routine iterative algorithm. If $\widehat{\theta}_{[t]}(u,\mathcal{B})$ and $\widehat{\theta}^{(1)}_{[t]}(u,\mathcal{B})$ represents the estimates at the tth iteration, then we update to $\widehat{\theta}_{[t+1]}(u,\mathcal{B})$ and $\widehat{\theta}^{(1)}_{[t+1]}(u,\mathcal{B})$ by solving $\mathbf{a}(u)$ from the equation

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_h(U_{ij}-u)\mathbf{G}_{ij}(u,h)\cdot w_i\mathcal{L}_{ij}(\boldsymbol{\eta}_i,\mathcal{B})|_{\boldsymbol{\eta}_i=\widehat{\boldsymbol{\eta}}_{i(j)[t]}(\mathcal{B},u)},$$

  where $\widehat{\boldsymbol{\eta}}_{i(j)[t]}(\mathcal{B},u) = (\widehat{\theta}_{[t]}(U_{i1},\mathcal{B}),\ldots,\widehat{\theta}_{[t]}(U_{i,j-1},\mathcal{B}), a_0(u)+a_1(u)(U_{ij}-u)/h, \widehat{\theta}_{[t]}(U_{i,j+1},\mathcal{B}),\ldots,\widehat{\theta}_{[t]}(U_{im},\mathcal{B})), j = 1,\ldots,m_i; t = 1,2,\ldots$. At convergence, with a given $\mathcal{B}$ we have the kernel estimator $\widehat{\theta}(u,\mathcal{B})$ and its derivative $\widehat{\theta}^{(1)}(u,\mathcal{B})$ for any u in the selected grid.
- **Stage 2**. Given the estimate $\widehat{\theta}(u;\mathcal{B}) = \widehat{a}_0(u)$ and $\widehat{a}_1(u)$ for points u in the selected grid, find the estimate of $\mathcal{B}$ by solving

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\mathbf{W}_i\cdot\mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_i,\mathcal{B})|_{\boldsymbol{\eta}_i=(\widehat{\theta}(U_{i1},\mathcal{B}),\ldots,\widehat{\theta}(U_{im},\mathcal{B}))}.$$

Repeat stages 1 and 2 until convergence. Denote by $\widehat{\mathcal{B}}_p$ the resulting profile kernel estimator for $\mathcal{B}$.

3.2. Asymptotic distributions

Analogous to Lin and Carroll (2001a,b) and Wang (2003), we take $m_i \equiv m$ for ease of exposition. Covariates $\mathbf{x}_i$ and $\mathbf{z}_i$ are allowed to be correlated, while the triples $(\mathbf{Y}_i,\mathbf{x}_i,\mathbf{z}_i), i = 1,\ldots,n$, are assumed independently identically distributed. For $i = 1,\ldots,n, j,k = 1,\ldots,m$, let $\mathcal{L}_{ijk}(\boldsymbol{\eta}_i,\mathcal{B}) = \partial^2\mathcal{L}_i(\boldsymbol{\eta}_i,\mathcal{B})/\partial\eta_{ij}\partial\eta_{ik}$,

$\mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_i, \mathcal{B}) = \partial^2 \mathcal{L}_i(\boldsymbol{\eta}, \mathcal{B})/\partial \eta_{ij} \partial \mathcal{B}$, and $\mathcal{L}_{i\mathcal{B}\mathcal{B}}(\boldsymbol{\eta}_i, \mathcal{B}) = \partial^2 \mathcal{L}_i(\boldsymbol{\eta}_i, \mathcal{B})/\partial \mathcal{B} \partial \mathcal{B}^T$. Let $\mathcal{B}_0 = (\boldsymbol{\alpha}_0^T, \boldsymbol{\beta}_0^T, \boldsymbol{\phi}_0^T)^T$ be the true value of parameter $\mathcal{B}$ and $\theta_0(\cdot)$ be the true function form of $\theta(\cdot)$. Take $U_{ij0} = \mathbf{z}_{ij}^T \boldsymbol{\alpha}_0$ and $\boldsymbol{\eta}_{i0} = (U_{i10}, \ldots, U_{im0})^T$. Let $\boldsymbol{\epsilon}_{ij}^{\#}(\theta, \mathcal{B}) = \mathbf{W}_i \cdot \mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_i, \mathcal{B}) + \sum_{k=1}^m \mathbf{W}_i \cdot \mathcal{L}_{ijk}(\boldsymbol{\eta}_i, \mathcal{B}) \cdot \boldsymbol{\theta}_{\mathcal{B}}(U_{ik}, \mathcal{B})$, where $\boldsymbol{\theta}_{\mathcal{B}}(u, \mathcal{B}_0)$ is the solution to

$$\mathbf{0} = \sum_{j=1}^m f_j(u) E\{\boldsymbol{\epsilon}_{ij}^{\#}(\theta_0, \mathcal{B}_0) | U_{ij0} = u\} \tag{3.2}$$

for a given u, where $f_j(u)$ is the marginal density of U_{ij0}.

Let $\Omega(u) = \sum_{j=1}^m f_j(u) E\{\mathcal{L}_{ijj}(\boldsymbol{\eta}_{i0}, \mathcal{B}_0) | U_{ij0} = u\}$. For a function $h(\cdot)$ and a point u, take

$$\Lambda(h, u) = \sum_{j=1}^m \sum_{k \neq j} f_j(u) \frac{E\{\mathcal{L}_{ijk}(\boldsymbol{\eta}_{i0}, \mathcal{B}_0) h(U_{ik0}) | U_{ij0} = u\}}{\Omega(u)},$$

and let function $b(u)$ be the solution to

$$b(u) = \theta_0^{(2)}(u) - \Lambda(b, u).$$

Adapting the proof in Lin and Carroll (2006), we establish the asymptotic distribution of the kernel estimator $\widehat{\theta}(\cdot)$.

Theorem 2. *Let $\lambda_1 = \int u^2 K(u) du$ and $\lambda_2 = \int K^2(u) du$. Suppose the bandwidth sequence satisfies $nh^2 \to \infty$ and $nh^6 \to 0$. Then, under regularity conditions similar to those of* Lin and Carroll (2006)*, we have*

$$\sqrt{nh} \left\{ \widehat{\theta}(u) - \theta_0(u) - \frac{1}{2} h^2 \lambda_1 b(u) \right\} \to_d N \left\{ 0, \frac{\lambda_2}{\Omega^2(u)} \sum_{j=1}^m E(D_{jj} | U_{ij0} = u) f_j(u) \right\},$$

where D_{jj} is the jth diagonal element of $\text{cov}(\boldsymbol{\epsilon}_i)$*, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im})^T$ with $\epsilon_{ij} = w_i \mathcal{L}_{ij}(\boldsymbol{\eta}_{i0}, \mathcal{B}_0)$.*

In the sequel, we establish the asymptotic distribution of the kernel profile estimator $\widehat{\mathcal{B}}_p$. Define $\mathcal{F}_1 = E\{\mathbf{W}_i \cdot \mathcal{L}_{i\mathcal{B}\mathcal{B}}(\boldsymbol{\eta}_{i0}, \mathcal{B}_0)\}, \mathcal{F}_2 = E\{\sum_{j=1}^m \mathbf{W}_i \cdot \mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_{i0}, \mathcal{B}_0) \boldsymbol{\theta}_{\mathcal{B}}^T(U_{ij}, \mathcal{B}_0)\}$, and $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$. Let

$$\mathcal{V} = \text{cov}\{\mathbf{W}_i \cdot \mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_0, \mathcal{B}_0) + \sum_{j=1}^m \mathbf{W}_i \cdot \mathcal{L}_{ij}(\boldsymbol{\eta}_0, \mathcal{B}_0) \cdot \boldsymbol{\theta}_{\mathcal{B}}(U_{ij0}, \mathcal{B}_0)\}.$$

In Appendix 2 we outline the proof of the following result.

Theorem 3. *Suppose that the bandwidth $h \propto n^{-c}$ with $1/5 \le c \le 1/3$. Then under the conditions of Theorems* 1 *and* 2*, we have, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\mathcal{B}}_p - \mathcal{B}_0) \to_d N(\mathbf{0}, \mathcal{F}^{-1}\mathcal{V}\mathcal{F}^{-1T}).$$

Inferences about parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\phi}$ can be based on Theorem 3, where $\mathcal{F}$ and $\mathcal{V}$ are replaced by consistent estimates in which the associated terms may be substituted by the empirical counterparts. Appendix 3 lists detailed expressions of the relevant derivatives. However, direct implementation of these estimates of $\mathcal{F}$ and $\mathcal{V}$ is complicated. Alternatively, one may apply the easily implemented bootstrap method for a variance estimate. This practice is widely invoked in settings with semiparametric models, see Lin and Carroll (2001a,b), Liang et al. (2004), and Wang, Carroll, and Lin (2005), for example.

4. Numerical Studies

4.1. Empirical assessment

We conducted simulation studies to evaluate the performance of the proposed method under two response models. In the first scenario we focused on pairwise association with higher order association being constrained as 0, while in the second scenario we included third order associations. To be specific, for $i = 1, \ldots, n$, we generated binary vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^T$ from the joint probability function

$$f(y_{i1}, \ldots, y_{im}) = \prod_{j=1}^{m} \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}} \Bigg\{ 1 + \sum_{j<k} \rho_{ijk} \cdot \frac{y_{ij}-\mu_{ij}}{\sqrt{v_{ijj}}} \cdot \frac{y_{ik}-\mu_{ik}}{\sqrt{v_{ikk}}} + \sum_{j<k<l} \xi \cdot \frac{y_{ij}-\mu_{ij}}{\sqrt{v_{ijj}}} \cdot \frac{y_{ik}-\mu_{ik}}{\sqrt{v_{ikk}}} \cdot \frac{y_{il}-\mu_{il}}{\sqrt{v_{ill}}} \Bigg\}, \qquad (4.1)$$

where ρ_{ijk} is the pairwise correlation coefficient of Y_{ij} and Y_{ik}, given by $\rho_{ijk} = (\mu_{ijk} - \mu_{ij}\mu_{ik})/\sqrt{v_{ijj}v_{ikk}}$, and $v_{ijj} = \mu_{ij}(1-\mu_{ij})$ is the marginal variance for Y_{ij}. A common third order association measure ξ was assumed among triples (Y_{ij}, Y_{ik}, Y_{il}). It was constrained to be 0 in Scenario I and set as 0.1 in Scenario II, featuring, respectively, a case that the model used to fit data coincides with or differs from the model of generating data.

The mean and association structures were respectively specified as

$$\text{logit}\ \mu_{ij} = \beta x_{ij} + \theta(\alpha_1 z_{i1} + \alpha_2 z_{i2} + \alpha_3 z_{i3}),$$

$$\log \psi_{ijk} = \phi,$$

where we took $\theta(t) = \sin\{\pi(t - 1.355\sqrt{3}/6)/(1.645\sqrt{3}/3)\}$, as considered in Carroll et al. (1997). Covariates x_{ij} were generated independently according to

Table 1. Empirical assessment of the performance of the proposed method with complete data.

		Bias					SE					MSE				
	ψ_{ijk}	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ
I[a]	0.5	0.015	-0.036	-0.021	-0.026	0.112	0.203	0.188	0.173	0.166	0.149	0.203	0.189	0.174	0.166	0.162
	1.0	0.001	-0.050	-0.023	-0.041	-0.007	0.204	0.223	0.208	0.183	0.197	0.204	0.225	0.208	0.185	0.197
	2.0	0.014	-0.091	-0.067	-0.030	0.110	0.192	0.292	0.292	0.183	0.256	0.193	0.300	0.296	0.184	0.269
II	0.5	0.024	0.014	-0.065	-0.041	0.195	0.203	0.186	0.179	0.183	0.152	0.204	0.187	0.183	0.184	0.190
	1.0	0.041	0.018	-0.062	-0.046	0.033	0.224	0.179	0.189	0.173	0.184	0.226	0.179	0.192	0.175	0.186
	2.0	0.082	-0.011	-0.044	-0.031	0.137	0.211	0.170	0.193	0.174	0.251	0.217	0.170	0.195	0.175	0.269

a Scenarios I and II correspond to model (4.1) with $\xi = 0$ and 0.1, respectively.

the binomial distribution $Bin(1, 0.5)$, and covariates z_{ij} were generated independently from the uniform distribution $U[0, 1]$. We set $\beta = 0.3$ and $\alpha_1 = \alpha_2 = \alpha_3 = 1/\sqrt{3}$. The odds ratio ψ_{ijk} was set to be 1.0, 0.5 and 2.0 to reflect varying strengths of association.

We considered a setting with $m = 4$ and $n = 100$. Two hundred simulations were run for each parameter configuration. In the simulation result tables, we report the average biases (Bias) of the differences between the true values and the estimates, the empirical standard errors (SE), and the mean squared errors (MSE) for the mean and association parameters.

First, we assess how the proposed method performed when data were complete. The simulation results are summarized in Table 1. It is not surprising that finite sample biases for parameter β tended to be the smallest. In Scenario I the estimates for association parameter ϕ had larger finite sample biases when association existed (i.e., $\psi_{ijk} \neq 1.0$) than when it did not. It appears that the standard errors and mean squared errors for the estimates of $\boldsymbol{\alpha}$ and ϕ increased as the value of ψ_{ijk} increased; while this trend did not seem to exist for the estimates of parameter β. In Scenario II, the model used to fit the data ignores the third order association existing in the model of generating data. Yet the proposed pairwise likelihood method still seems to produce fairly reasonable estimates.

Next, we evaluate the performance of the proposed method when missing observations were present. The missing data indicator was generated independently from the logistic regression model

$$\text{logit}P(R_{ij} = 1|\mathbf{y}_i) = \gamma_0 + \gamma_1 y_{i,j-1} + \gamma_2 y_{ij}.$$

We took $\gamma_0 = 0.5$ and $\gamma_1 = 0.1$, and let γ_2 be 0.5, 0, -0.5, yielding varying missingness proportions that varied roughly between 25% and 50%.

To see the possible impact of different choices of the weight matrix $\mathbf{W}_i$, we considered the case with or without weights. That is, in implementing the method discussed in Section 3, $\mathbf{W}_i$ was taken as the identity matrix $\mathbf{I}_{5\times5}$ or

Table 2. Empirical assessment of the performance of the proposed method with incomplete data: No weight adjustments for varying numbers of missing observations.

			Bias					SE					MSE				
	γ_2	ψ_{ijk}	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ
I[a]	0.0	0.5	0.008	-0.067	-0.025	-0.067	0.108	0.219	0.256	0.222	0.247	0.214	0.219	0.260	0.222	0.251	0.225
		1.0	-0.005	-0.042	-0.048	-0.054	0.001	0.205	0.257	0.237	0.196	0.268	0.205	0.259	0.239	0.199	0.268
		2.0	0.011	-0.140	-0.107	-0.070	0.217	0.211	0.378	0.344	0.266	0.347	0.211	0.397	0.355	0.271	0.394
	0.5	0.5	0.020	-0.058	-0.060	-0.057	0.119	0.208	0.277	0.249	0.233	0.221	0.208	0.281	0.253	0.236	0.235
		1.0	0.007	-0.067	-0.065	-0.027	-0.020	0.198	0.234	0.248	0.240	0.276	0.198	0.239	0.252	0.241	0.277
		2.0	-0.008	-0.163	-0.116	-0.046	0.170	0.228	0.406	0.314	0.265	0.302	0.228	0.432	0.328	0.267	0.331
	-0.5	0.5	0.012	-0.087	-0.051	-0.043	0.082	0.209	0.308	0.236	0.218	0.251	0.209	0.316	0.238	0.219	0.257
		1.0	0.001	-0.077	-0.055	-0.029	0.008	0.206	0.267	0.257	0.198	0.266	0.206	0.272	0.260	0.199	0.267
		2.0	0.008	-0.099	-0.101	-0.028	0.198	0.210	0.299	0.329	0.215	0.313	0.210	0.308	0.339	0.215	0.352
II	0.0	0.5	0.015	0.007	-0.088	-0.020	0.208	0.272	0.175	0.204	0.192	0.232	0.272	0.175	0.212	0.192	0.275
		1.0	-0.008	0.010	-0.091	-0.048	0.034	0.246	0.205	0.209	0.229	0.282	0.246	0.205	0.217	0.231	0.283
		2.0	0.022	0.011	-0.117	-0.040	0.129	0.260	0.194	0.252	0.228	0.315	0.260	0.194	0.265	0.230	0.332
	0.5	0.5	0.045	-0.006	-0.070	-0.071	0.229	0.242	0.178	0.225	0.282	0.290	0.244	0.178	0.230	0.287	0.342
		1.0	0.004	-0.006	-0.094	-0.018	0.011	0.269	0.195	0.210	0.214	0.260	0.269	0.195	0.218	0.214	0.260
		2.0	0.004	-0.002	-0.061	-0.031	0.155	0.277	0.178	0.184	0.199	0.298	0.277	0.178	0.188	0.200	0.322
	-0.5	0.5	-0.003	0.024	-0.111	-0.040	0.189	0.257	0.192	0.217	0.221	0.246	0.257	0.193	0.229	0.222	0.281
		1.0	0.025	-0.004	-0.089	-0.017	0.039	0.238	0.193	0.207	0.197	0.278	0.239	0.193	0.215	0.197	0.279
		2.0	-0.003	-0.025	-0.096	-0.013	0.146	0.269	0.193	0.221	0.243	0.307	0.269	0.193	0.230	0.243	0.328

[a]Scenarios I and II correspond to model (4.1) with $\xi = 0$ and 0.1, respectively.

diagonal matrix $w_i\mathbf{I}_{5\times 5}$ with $w_i = 1/(O_i - 1)$. Table 2 displays the simulation results corresponding to the case without weights. The same patterns as those in Table 1 are observed here. Finite sample biases for the β pareameter tended to be the smallest. Estimates of the ϕ parameter seemed to involve smaller biases for $\phi = 0$ than for $\phi \neq 0$. Comparing the results in Scenarios I and II, we see that the impact of third order association on the performance of the pairwise likelihood appears to agree with the expectation. The pairwise likelihood method seems fairly robust against misspecification of third order associations.

In Table 3 we report on the simulation results for the case with weights incorporated. The impact of adding weights seems to be more visible on estimation of mean parameters than on estimation of association parameter ϕ. Again, the trends revealed by Scenarios I and II are fairly comparable with those in Tables 1 and 2. In summary, the proposed method performed fairly satisfactorily under various settings, with or without missing observations, with different missing data proportions and with varying strengths of association measures.

4.2. An application

In this subsection we apply the proposed method to analyze a family data set of the Genetic Analysis Workshop (GAW13) arising from the Framingham Heart Study. The Framingham Heart Study is an ongoing prospective study of

Table 3. Empirical Assessment of the Performance of the Proposed Method with Incomplete Data: Inclusion of Weight Adjustments for Varying Numbers of Missing Observations

			Bias					SE					MSE				
	γ_2	ψ_{ijk}	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ	β	α_1	α_2	α_3	ϕ
I[a]	0.0	0.5	-0.007	0.026	-0.080	-0.046	0.108	0.277	0.183	0.201	0.181	0.244	0.277	0.184	0.207	0.184	0.255
		1.0	0.017	0.011	-0.076	-0.050	-0.020	0.281	0.188	0.230	0.191	0.322	0.281	0.188	0.235	0.194	0.323
		2.0	0.009	-0.003	-0.105	-0.028	0.114	0.277	0.186	0.221	0.250	0.335	0.277	0.186	0.232	0.251	0.348
	0.5	0.5	0.016	-0.013	-0.056	-0.040	0.087	0.252	0.203	0.202	0.200	0.271	0.252	0.203	0.205	0.202	0.278
		1.0	0.043	0.008	-0.114	-0.027	-0.004	0.240	0.188	0.236	0.220	0.272	0.242	0.188	0.249	0.221	0.272
		2.0	-0.005	0.007	-0.108	-0.059	0.187	0.283	0.194	0.257	0.257	0.320	0.283	0.194	0.269	0.260	0.355
	-0.5	0.5	0.006	0.006	-0.076	-0.060	0.064	0.261	0.180	0.238	0.229	0.260	0.261	0.180	0.244	0.233	0.265
		1.0	0.045	0.018	-0.082	-0.032	-0.011	0.267	0.185	0.195	0.176	0.286	0.269	0.186	0.202	0.177	0.286
		2.0	0.035	-0.019	-0.085	0.002	0.179	0.281	0.194	0.198	0.185	0.333	0.282	0.195	0.205	0.185	0.365
II	0.0	0.5	0.002	-0.126	-0.085	0.016	0.243	0.232	0.287	0.258	0.232	0.273	0.232	0.302	0.265	0.232	0.332
		1.0	0.006	-0.147	-0.045	0.010	0.035	0.261	0.272	0.249	0.226	0.273	0.261	0.293	0.251	0.226	0.274
		2.0	0.011	-0.145	-0.038	0.008	0.082	0.253	0.261	0.234	0.239	0.307	0.253	0.282	0.236	0.239	0.314
	0.5	0.5	0.003	-0.174	-0.057	0.028	0.222	0.263	0.296	0.244	0.230	0.255	0.263	0.326	0.247	0.231	0.305
		1.0	0.002	-0.145	-0.055	0.020	0.039	0.249	0.267	0.230	0.244	0.271	0.249	0.288	0.233	0.245	0.273
		2.0	0.013	-0.123	-0.062	0.006	0.125	0.264	0.270	0.239	0.239	0.299	0.264	0.286	0.243	0.239	0.314
	-0.5	0.5	-0.017	-0.155	-0.057	0.021	0.239	0.233	0.278	0.246	0.238	0.275	0.233	0.302	0.249	0.238	0.332
		1.0	-0.011	-0.167	-0.077	0.042	0.015	0.261	0.279	0.241	0.249	0.266	0.261	0.307	0.247	0.251	0.266
		2.0	0.027	-0.144	-0.02	-0.018	0.085	0.296	0.265	0.243	0.245	0.297	0.297	0.285	0.244	0.246	0.304

[a]Scenarios I and II correspond to model (4.1) with $\xi = 0$ and 0.1, respectively.

risk factors for cardiovascular disease (CVD). The objective of the Framingham Heart Study was to identify common factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of participants who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke. The family data from the Framingham Heart Study were collected across two cohorts. The original Framingham participants were between 29-62 years of age at the start of the study, and data on these participants were available for 21 examination periods at 2-year intervals between 1948-1988. A second cohort, the Framingham Offspring Study, composed of children of members of the first cohort, was followed from 1971-1991, with five examinations over this 20-year period. A full description of the GAW13 Framingham Heart Study data set is provided by Cupples et al. (2003).

There were 326 families of 1672 individuals in the Framingham Offspring Cohort Data provided for GAW13. For illustration we apply the method discussed in Section 3 to a subset to ease computation. There were 126 families which had more than four individuals. We include the first four individuals for those families in our analysis. Baseline measurements are used in the analysis here.

High blood pressure is an important risk factor for cardiovascular disease and is a leading cause of mortality in industrialized countries. However, it is a complex disorder that results from environmental and genetic factors and their

Table 4. Analyses of a family data set from the framingham heart study.

	Gender	Age	HDL	BMI	Association
Est.	0.915	0.642	0.369	0.672	1.679
SE	0.260	0.209	0.142	0.166	0.425

interactions (Kraft et al. (2003)). It is of interest to study what risk factors may be associated with blood pressure and whether or not individuals within the same family are correlated in terms of responses. The covariates of interest include age, gender, high density lipoprotein (HDL), and body mass index (BMI) (BMI=weight (kg)/height^2 (m^2)). Let $Y_{ij} = 1$ if subject j in family i has high blood pressure, and $Y_{ij} = 0$ otherwise.

We consider a semiparametric regression model for the mean response, specified as

$$\text{logit } \mu_{ij} = \beta x_{ij} + \theta(\alpha_1 z_{ij1} + \alpha_2 z_{ij2} + \alpha_3 z_{ij3}),$$

where x_{ij} is gender, taking value 1 for male and 0 otherwise, z_{ij1} is age, z_{ij2} is HDL, and z_{ij2} is BMI. z_{ij1}, z_{ij2} and z_{ij3} are standardized as $\Phi((z_{ijk} - \bar{z}_{..k})/s_{..k})$, where $\bar{z}_{..k}$ and $s_{..k}$ represent the sample mean and standard deviation of z'_{ijk}s, respectively, $k = 1, 2, 3$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Exchangeable association structure is modeled here with $\log \psi_{ijk} = \phi, j \neq k$.

To conduct estimation we take the standard normal density as the kernel, and a data-driven bandwidth h is used. Table 4 reports the parameter estimates and their bootstrap standard errors, and Figure 1 shows the estimates of single index curve $\theta(\cdot)$ evaluated for the female and male data. The estimates of the single index curve $\theta(\cdot)$ show clear nonlinear curvatures, and this suggests that the data may not be fitted well by an ordinary logistic regression model. Inclusion of the nonlinear single index term $\theta(\cdot)$ in the model allows more flexibility to capture the nonlinear trend of the data. The interpretation of nonparametric covariate effects $\boldsymbol{\alpha}$ is not as transparent as that for parametric coefficients. In principle, as suggested in Carroll et al. (1997), non-zero estimates of $\boldsymbol{\alpha}$ often indicate significant effects. The estimates of the $\boldsymbol{\alpha}$ parameters suggest that age, HDL, and BMI are “significant” predictors for blood pressure. The analysis finds evidence for a positive association of high blood pressure existing among family members.

5. Discussion

Modeling association parameters simultaneously, with regression coefficients for marginal means, has been advocated for use on the grounds of improved

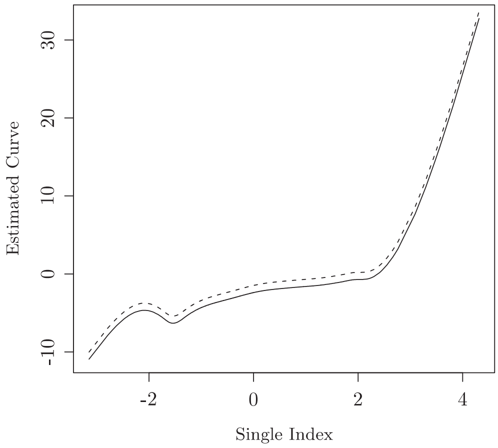

Figure 1. Estimated nonlinear curves for the family data from the Framingham Heart Study. Solid curve is the estimate of logit$\{P$(high blood pressure)$\}$ for females, and the dotted curve is the estimate of logit$\{P$(high blood pressure)$\}$ for males.

efficiency for parameter estimation under a parametric setup (Liang, Zeger, and Qaqish (1992)). Relatively little attention has been directed to semiparametric settings. Recently, with correlated binary data, Yi, He, and Liang (2009, 2010) developed simultaneous inference strategies for mean and association parameters under generalized partially linear single-index models. However, those methods mainly apply to complete data as they are based on the GEE formulation. In the presence of missing observations, those methods may not yield valid inference unless proper weights are included to adjust for the missingness effects. In this paper, we consider the same model setup for the response process, and exploit an inference method that is flexible to handle either complete or incomplete data under a unified framework. An appealing feature of the proposed method is that the missing data process is left unspecified when missing observations are present, yet model assumptions for the response process are kept minimal. The simulation study demonstrates reasonable performance of the proposed method.

The method we describe here has applications in a wide variety of settings. They can also be generalized to accommodating data with more complex association structures. For example, in many situations longitudinal data arise in clusters for which both a cross-sectional and a longitudinal correlation exist, and interest may reside on the strengths of both types of association (Yi and Cook (2002a,b)). To handle such data, a single model is often inadequate to facilitate complex association structures, but different types of regression models are normally required. The proposed method can be modified to accommodate such cases.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Boston University and NHLBI for providing the data set from the Framingham Heart Study (No. N01-HC-25195) in the illustration. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI.

Appendix 1: Proof of Theorem 1

In the following derivations, we let $f(\cdot)$ and $f(\cdot|\cdot)$ denote the probability function and the conditional probability function for the random vectors that are indicated by the arguments. Dependence on the corresponding parameters is suppressed in the notation. Write the realization vector $\mathbf{r}_i$ of the missing data indicator vector $\mathbf{R}_i$ as $\mathbf{r}_i = (r_{ij}, r_{ik}, \mathbf{r}_{(i;jk)})$, where $\mathbf{r}_{(i;jk)}$ denotes the subvector of $\mathbf{r}_i$ with components r_{ij} and r_{ik} excluded. Similar notation $\mathbf{y}_{(i;jk)}$ is defined for the response subvector. Let $|\mathbf{r}_{(i;jk)}|$ denote the sum of the elements of the subvector $\mathbf{r}_{(i;jk)}$.

The proof is similar to that of Yi, Zeng and Cook (2010), except that a nonlinear unknown vector $\boldsymbol{\eta}_i$ is involved here. It suffices to show that $E\{\sum_{j<k} [1/(O_i-1)](\partial/\partial\mathcal{B})\log(L_{ijk}^{R_{ij}R_{ik}})\} = \mathbf{0}$. The proof for $E[S_{ij}] = 0$ follows analogously. Indeed,

$$
\begin{aligned}
&E\left\{\sum_{j<k}\frac{1}{O_i-1}\cdot\frac{\partial}{\partial\mathcal{B}}\log(L_{ijk}^{R_{ij}R_{ik}})\right\}\\
&=\sum_{j<k}E\left\{\frac{R_{ij}R_{ik}}{\sum_s R_{is}-1}\cdot\frac{\partial\log f(y_{ij},y_{ik}|\mathbf{x}_i,\mathbf{z}_i)}{\partial\mathcal{B}}\right\}\\
&=\sum_{j<k}\sum_{\mathbf{r}_i}\sum_{\mathbf{y}_i}\frac{r_{ij}r_{ik}}{\sum_s r_{is}-1}\cdot\frac{\partial\log f(y_{ij},y_{ik}|\mathbf{x}_i,\mathbf{z}_i)}{\partial\mathcal{B}}\cdot f(\mathbf{r}_i,\mathbf{y}_i|\mathbf{x}_i,\mathbf{z}_i)\\
&=\sum_{j<k}\sum_{\mathbf{r}_i}\sum_{y_{ij},y_{ik}}\sum_{\mathbf{y}_{(i;jk)}}\frac{r_{ij}r_{ik}}{\sum_s r_{is}-1}\cdot\frac{\partial\log f(y_{ij},y_{ik}|\mathbf{x}_i,\mathbf{z}_i)}{\partial\mathcal{B}}\cdot f(\mathbf{r}_i,y_{ij},y_{ik},\mathbf{y}_{(i;jk)}|\mathbf{x}_i,\mathbf{z}_i)\\
&=\sum_{j<k}\sum_{y_{ij},y_{ik}}\sum_{\mathbf{r}_i}\frac{r_{ij}r_{ik}}{\sum_s r_{is}-1}\cdot\frac{\partial\log f(y_{ij},y_{ik}|\mathbf{x}_i,\mathbf{z}_i)}{\partial\mathcal{B}}\cdot f(\mathbf{r}_i,y_{ij},y_{ik}|\mathbf{x}_i,\mathbf{z}_i)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j<k} \sum_{y_{ij},y_{ik}} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \cdot \frac{\partial \log f(y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i)}{\partial \mathcal{B}} \\
&\quad \cdot f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}, y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i) \\
&= \sum_{j<k} \sum_{y_{ij},y_{ik}} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \cdot \frac{\partial f(y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i)}{\partial \mathcal{B}} \\
&\quad \cdot f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i) \\
&= \sum_{j<k} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \\
&\quad \cdot \sum_{y_{ij},y_{ik}} \left\{ \frac{\partial f(y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i)}{\partial \mathcal{B}} \cdot f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i) \right\} \\
&= \sum_{j<k} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \\
&\quad \cdot \frac{\partial}{\partial \mathcal{B}} \left\{ \sum_{y_{ij},y_{ik}} f(y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i) \cdot f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i) \right\} \\
&= \sum_{j<k} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \cdot \frac{\partial}{\partial \mathcal{B}} \left\{ \sum_{y_{ij},y_{ik}} f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}, y_{ij}, y_{ik}|\mathbf{x}_i, \mathbf{z}_i) \right\} \\
&= \sum_{j<k} \sum_{\mathbf{r}_{(i;jk)}} \frac{1}{|\mathbf{r}_{(i;jk)}|+1} \cdot \frac{\partial}{\partial \mathcal{B}} \left\{ f(r_{ij} = r_{ik} = 1, \mathbf{r}_{(i;jk)}|\mathbf{x}_i, \mathbf{z}_i) \right\} \\
&= \mathbf{0}
\end{aligned}
$$

where, in the fourth last step, we impose the assumption that $f(\mathbf{r}_i|y_{ij}, y_{ik}, \mathbf{x}_i, \mathbf{z}_i)$ does not depend on the response parameter $\mathcal{B}$, and in the last step we apply the assumption that the distribution of $f(\mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i)$ is free of the response parameter $\mathcal{B}$.

Appendix 2: Proof of Theorem 3

The proof of this theorem shares the same spirit as A.4 concerning the profiling estimator discussed in Lin and Carroll (2006). However, the development here cannot simply be treated as a particular application of the results of Lin and Carroll (2006). There are a couple of important features that distinguish the current development from settings considered in Lin and Carroll (2006) First, the nonlinear component in the model is more complex here with additional parameters $\boldsymbol{\alpha}$ to be estimated. Secondly, extra association parameters $\boldsymbol{\phi}$ are required to be estimated. Finally, and most importantly, the current development is flexible with missingness accommodated.

In the following derivations, $U_{ij0} = \mathbf{z}_{ij}^T\boldsymbol{\alpha}_0$ plays a similar role to that of Z_{ij} in Lin and Carroll (2006). Let $f_{jk}(u,v)$ be the joint density of U_{ij0} and U_{ik0}, $j \neq k$. Write $\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(u,\mathcal{B}) = \partial\widehat{\theta}(u,\mathcal{B})/\partial\mathcal{B}$. Denote by $\boldsymbol{\theta}_{\mathcal{B}\mathcal{B}}(u,\mathcal{B})$ the limit of $\partial^2\widehat{\theta}(u,\mathcal{B})/\partial\mathcal{B}\partial\mathcal{B}^T$ as $n \to \infty$. Adapting the arguments in A.3 and A.4 of Lin and Carroll (2006), with U_{ij0} and u replacing Z_{ij} and z respectively, we can show that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(u,\mathcal{B}_0) &= \boldsymbol{\theta}_{\mathcal{B}}(u,\mathcal{B}_0) + o_p(1),\\ \widehat{\boldsymbol{\theta}}_{\mathcal{B}\mathcal{B}}(u,\mathcal{B}_0) &= \boldsymbol{\theta}_{\mathcal{B}\mathcal{B}}(u,\mathcal{B}_0) + o_p(1),\end{aligned} \tag{A.1}$$

i.e., $\boldsymbol{\theta}_{\mathcal{B}}(u,\mathcal{B}_0)$ defined by (3.2) can be viewed as the limit of $\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(u,\mathcal{B}_0)$ as $n \to \infty$.

Let $\mathbf{H}_j(u) = E\{\boldsymbol{\epsilon}_{ij}^{\#}(\theta_0,\mathcal{B}_0)|U_{ij0} = u\}$, then for any function $B(\cdot)$, we have

$$\begin{aligned}E\{\sum_{j=1}^m B(U_{ij0})\cdot\mathbf{H}_j(U_{ij0})\} &= \sum_{j=1}^m E\{B(U_{ij0})\cdot\mathbf{H}_j(U_{ij0})\}\\ &= \sum_{j=1}^m \int B(u)\cdot\mathbf{H}_j(u)f_j(u)du = \int B(u)\cdot\{\sum_{j=1}^m \mathbf{H}_j(u)f_j(u)\}du\\ &= \mathbf{0},\end{aligned} \tag{A.2}$$

where the last step is due to (3.2). If $\mathcal{F}_3 = E\{\sum_{j=1}^m\sum_{k=1}^m \mathbf{W}_i\cdot\mathcal{L}_{ijk}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\,\boldsymbol{\theta}_{\mathcal{B}}(U_{ij},\mathcal{B}_0)\boldsymbol{\theta}_{\mathcal{B}}^T(U_{ij},\mathcal{B}_0)\}$ then, by (A.2), we have $\mathcal{F}_2 + \mathcal{F}_3 = \mathbf{0}$, hence $\mathcal{F} = \mathcal{F}_1 + 2\mathcal{F}_2 + \mathcal{F}_3$.

Now we show the asymptotic distribution for the profile estimator $\widehat{\mathcal{B}}_p$. Let

$$\begin{aligned}\mathbf{A}_1(\widehat{\mathcal{B}}_p,\widehat{\theta}) &= n^{-1/2}\sum_{i=1}^n \mathbf{W}_i\cdot\mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta},\mathcal{B})|_{(\boldsymbol{\eta},\mathcal{B})=((\widehat{\theta}(U_{i1},\widehat{\mathcal{B}}_p),\ldots,\widehat{\theta}(U_{im},\widehat{\mathcal{B}}_p)),\widehat{\mathcal{B}}_p)},\\ \mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta}) &= n^{-1/2}\sum_{i=1}^n\sum_{j=1}^m \mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta},\mathcal{B})|_{(\boldsymbol{\eta},\mathcal{B})=((\widehat{\theta}(U_{i1},\widehat{\mathcal{B}}_p),\ldots,\widehat{\theta}(U_{im},\widehat{\mathcal{B}}_p)),\widehat{\mathcal{B}}_p)}\cdot\widehat{\theta}_{\mathcal{B}}(U_{ij},\widehat{\mathcal{B}}_p).\end{aligned}$$

Then the profile estimator $\widehat{\mathcal{B}}_p$ sovles the equation

$$\mathbf{0} = \mathbf{A}_1(\widehat{\mathcal{B}}_p,\widehat{\theta}) + \mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta}).$$

To sort out the leading terms from the higher order terms, we apply Taylor series expansions to $\mathbf{A}_1(\widehat{\mathcal{B}}_p,\widehat{\theta})$ and $\mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta})$. Specifically, we expand $\mathbf{A}_1(\widehat{\mathcal{B}}_p,\widehat{\theta})$ around the true $(\mathcal{B}_0,\theta_0)$ and use $\widehat{\theta}(U_{ij},\widehat{\mathcal{B}}_p) - \theta_0(U_{ij0}) = \widehat{\theta}(U_{ij},\mathcal{B}_0) - \theta_0(U_{ij0}) + o_p(1)$:

$$\mathbf{A}_1(\widehat{\mathcal{B}}_p,\widehat{\theta}) = n^{-1/2}\sum_{i=1}^n \mathbf{W}_i\cdot\mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)$$

$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\{\widehat{\theta}(U_{ij},\mathcal{B}_0)-\theta_0(U_{ij0})\}$$
$$+(\mathcal{F}_1+\mathcal{F}_2)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)+o_p(1). \quad \text{(A.3)}$$

For $\mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta})$, we first treat it as a function of $\mathcal{B}$ and expand it around $\mathcal{B}_0$ by using (A.1):

$$\mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta})=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\widehat{\theta}(U_{i1},\mathcal{B}_0),\ldots,\widehat{\theta}(U_{im},\mathcal{B}_0),\mathcal{B}_0)\cdot\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(U_{ij},\mathcal{B}_0)$$
$$+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_i,\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}^T(U_{ij},\mathcal{B}_0)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)$$
$$+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_i,\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}\mathcal{B}}^T(U_{ij},\mathcal{B}_0)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)$$
$$+n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ijk}(\boldsymbol{\eta}_i,\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}(U_{ij},\mathcal{B}_0)\boldsymbol{\theta}_{\mathcal{B}}^T(U_{ik},\mathcal{B}_0)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)$$
$$+o_p(1).$$

It is easily seen that the second and last terms sum to $(\mathcal{F}_2+\mathcal{F}_3)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)+o_p(1)$. Adapting the arguments in Appendix 1, we can show that $E\{\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_i,\mathcal{B}_0)|U_{i10},\ldots,U_{im0}\}=\mathbf{0}$, hence the third term is $o_p(1)$. Now further decomposing the first term around $(\boldsymbol{\eta}_{i0},\mathcal{B}_0)$ leads to

$$\mathbf{A}_2(\widehat{\mathcal{B}}_p,\widehat{\theta})$$
$$=(\mathcal{F}_2+\mathcal{F}_3)n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)$$
$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ijk}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)\cdot\{\widehat{\theta}(U_{ik0},\mathcal{B}_0)-\theta_0(U_{ik0})\}$$
$$+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\{\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)-\boldsymbol{\theta}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)\}+o_p(1). \quad \text{(A.4)}$$

Let $\mathbf{P}_{ij}=\mathbf{W}_i\cdot\mathcal{L}_{ij\mathcal{B}}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)+\sum_{k=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ijk}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}(U_{ik0},\mathcal{B}_0)$, then combining (A.3) and (A.4) gives

$$-\mathcal{F}n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)$$
$$=n^{-1/2}\sum_{i=1}^{n}\{\mathbf{W}_i\cdot\mathcal{L}_{i\mathcal{B}}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)+\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\boldsymbol{\theta}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)\}$$

$$
\begin{aligned}
&+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{H}_j(U_{ij0})\cdot\{\widehat{\theta}(U_{ij0},\mathcal{B}_0)-\theta_0(U_{ij0})\}\\
&+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\{\mathbf{P}_{ij}-\mathbf{H}_j(U_{ij0})\}\cdot\{\widehat{\theta}(U_{ij0},\mathcal{B}_0)-\theta_0(U_{ij0})\}\\
&+n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{W}_i\cdot\mathcal{L}_{ij}(\boldsymbol{\eta}_{i0},\mathcal{B}_0)\cdot\{\widehat{\boldsymbol{\theta}}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)-\boldsymbol{\theta}_{\mathcal{B}}(U_{ij0},\mathcal{B}_0)\}\\
&+o_p(1). \qquad\qquad (\text{A.5})
\end{aligned}
$$

We can show the last three terms of (A.5) are all $o_p(1)$ by adapting the arguments in Lin and Carroll (2006). Therefore, by the Central Limit Theorem, the asymptotic distribution of $n^{1/2}(\widehat{\mathcal{B}}_p-\mathcal{B}_0)$ is $N(\mathbf{0},\mathcal{F}^{-1}\mathcal{V}\mathcal{F}^{-1T})$.

Appendix 3: Computation Details

Here we present the detailed expressions for the derivatives that may be used in the estimation and inferential procedures. For $j\neq k$, let $\boldsymbol{\delta}=(\mathcal{B}^T,\eta_{ij},\eta_{ik})^T$, $\boldsymbol{\zeta}=(\boldsymbol{\beta}^T,\boldsymbol{\alpha}^T,\eta_{ij},\eta_{ik})^T$, and let

$$
\begin{aligned}
L_{ijk}=P(Y_{ij}=y_{ij},Y_{ik}=y_{ik}|\mathbf{x}_i,\mathbf{z}_i)&=\mu_{ijk}^{y_{ij}y_{ik}}\cdot(\mu_{ij}-\mu_{ijk})^{y_{ij}(1-y_{ik})}\\
&\cdot(\mu_{ik}-\mu_{ijk})^{(1-y_{ij})y_{ik}}\cdot(1-\mu_{ij}-\mu_{ik}+\mu_{ijk})^{(1-y_{ij})(1-y_{ik})}
\end{aligned}
$$

be the pairwise likelihood.

First, we record the first derivatives:

$$
\begin{aligned}
\frac{\partial\log L_{ijk}}{\partial\boldsymbol{\delta}}&=\frac{y_{ij}y_{ik}}{\mu_{ijk}}\cdot\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}}+\frac{y_{ij}(1-y_{ik})}{\mu_{ij}-\mu_{ijk}}\cdot\left(\frac{\partial\mu_{ij}}{\partial\boldsymbol{\delta}}-\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}}\right)\\
&+\frac{(1-y_{ij})y_{ik}}{\mu_{ik}-\mu_{ijk}}\cdot\left(\frac{\partial\mu_{ik}}{\partial\boldsymbol{\delta}}-\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}}\right)\\
&-\frac{(1-y_{ij})(1-y_{ik})}{1-\mu_{ij}-\mu_{ik}+\mu_{ijk}}\cdot\left(\frac{\partial\mu_{ij}}{\partial\boldsymbol{\delta}}+\frac{\partial\mu_{ik}}{\partial\boldsymbol{\delta}}-\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}}\right).
\end{aligned}
$$

More specifically, the first derivatives of μ_{ij} and μ_{ijk} are given as follows.

(1) For the marginal probability, we have the derivatives:

$$
\begin{aligned}
&\frac{\partial\mu_{ij}}{\partial\boldsymbol{\beta}}=\mathbf{x}_{ij}g^{(1)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\quad \frac{\partial\mu_{ij}}{\partial\boldsymbol{\alpha}}=\mathbf{z}_{ij}\theta^{(1)}(\mathbf{z}_{ij}^T\boldsymbol{\alpha})g^{(1)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\\
&\frac{\partial\mu_{ij}}{\partial\boldsymbol{\phi}}=\mathbf{0},\quad \frac{\partial\mu_{ij}}{\partial\eta_{ij}}=g^{(1)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\ \text{ and } \frac{\partial\mu_{ij}}{\partial\eta_{ik}}=0,\ j\neq k.
\end{aligned}
$$

(2) For the pairwise probability, the derivatives are given by

$$\frac{\partial \mu_{ijk}}{\partial \boldsymbol{\zeta}} = \begin{cases} \frac{1}{2(\psi_{ijk}-1)}\left[\frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}} - \frac{1}{2\sqrt{b_{ijk}}}\frac{\partial b_{ijk}}{\partial \boldsymbol{\zeta}}\right], \psi_{ijk} \neq 1, \\ \frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}}\mu_{ik} + \mu_{ij}\frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}}, \qquad \psi_{ijk} = 1, \end{cases}$$

$$\frac{\partial \mu_{ijk}}{\partial \boldsymbol{\phi}} = \begin{cases} \frac{1}{(\psi_{ijk}-1)}\left[-\mu_{ijk}\frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}} + \frac{1}{2}\left(\frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}} - \frac{1}{2\sqrt{b_{ijk}}}\frac{\partial b_{ijk}}{\partial \boldsymbol{\phi}}\right)\right], \psi_{ijk} \neq 1, \\ \mathbf{0}, \qquad \psi_{ijk} = 1, \end{cases}$$

where

$$\frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}} = -(1-\psi_{ijk})\left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}} + \frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}}\right),$$

$$\frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}} = \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}}(\mu_{ij} + \mu_{ik}),$$

$$\frac{\partial b_{ijk}}{\partial \boldsymbol{\zeta}} = 2a_{ijk}\frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}} - 4\psi_{ijk}(1-\psi_{ijk})\left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}}\mu_{ik} + \mu_{ij}\frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}}\right),$$

$$\frac{\partial b_{ijk}}{\partial \boldsymbol{\phi}} = 2a_{ijk}\frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}} - 4(2\psi_{ijk}-1)\mu_{ij}\mu_{ik}\frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}}.$$

Secondly, we display the second derivatives:

$$\frac{\partial^2 \log L_{ijk}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}$$

$$= -\frac{y_{ij}y_{ik}}{\mu_{ijk}^2}\cdot\frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}}\cdot\frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}^T} + \frac{y_{ij}y_{ik}}{\mu_{ijk}}\cdot\frac{\partial^2 \mu_{ijk}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}$$

$$-\frac{y_{ij}(1-y_{ik})}{(\mu_{ij}-\mu_{ijk})^2}\cdot\left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\delta}} - \frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}}\right)\cdot\left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\delta}^T} - \frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}^T}\right)$$

$$+\frac{y_{ij}(1-y_{ik})}{\mu_{ij}-\mu_{ijk}}\cdot\left(\frac{\partial^2 \mu_{ij}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} - \frac{\partial^2 \mu_{ijk}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}\right)$$

$$-\frac{(1-y_{ij})y_{ik}}{(\mu_{ik}-\mu_{ijk})^2}\cdot\left(\frac{\partial \mu_{ik}}{\partial \boldsymbol{\delta}} - \frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}}\right)\cdot\left(\frac{\partial \mu_{ik}}{\partial \boldsymbol{\delta}^T} - \frac{\partial \mu_{ijk}}{\partial \boldsymbol{\delta}^T}\right)$$

$$+\frac{(1-y_{ij})y_{ik}}{\mu_{ik}-\mu_{ijk}}\cdot\left(\frac{\partial^2 \mu_{ik}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T} - \frac{\partial^2 \mu_{ijk}}{\partial \boldsymbol{\delta}\partial \boldsymbol{\delta}^T}\right)$$

$$-\frac{(1-y_{ij})(1-y_{ik})}{(1-\mu_{ij}-\mu_{ik}+\mu_{ijk})^2}\cdot\left(\frac{\partial\mu_{ij}}{\partial\boldsymbol{\delta}}+\frac{\partial\mu_{ik}}{\partial\boldsymbol{\delta}}-\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}}\right)\cdot\left(\frac{\partial\mu_{ij}}{\partial\boldsymbol{\delta}^T}+\frac{\partial\mu_{ik}}{\partial\boldsymbol{\delta}^T}-\frac{\partial\mu_{ijk}}{\partial\boldsymbol{\delta}^T}\right)$$
$$-\frac{(1-y_{ij})(1-y_{ik})}{1-\mu_{ij}-\mu_{ik}+\mu_{ijk}}\cdot\left(\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}+\frac{\partial^2\mu_{ik}}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}-\frac{\partial^2\mu_{ijk}}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T}\right).$$

More specifically, the second derivatives of μ_{ij} and μ_{ijk} are given as follows.

(a) The second derivatives of μ_{ij} are:

$$\begin{aligned}
\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} &= \mathbf{z}_{ij}\mathbf{z}_{ij}^T\theta^{(2)}(\mathbf{z}_{ij}^T\boldsymbol{\alpha})g^{(1)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha}))\\
&\quad+\mathbf{z}_{ij}\mathbf{z}_{ij}^T\{\theta^{(1)}(\mathbf{z}_{ij}^T\boldsymbol{\alpha})\}^2g^{(2)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\\
\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^T} &= \mathbf{z}_{ij}\mathbf{x}_{ij}^T\theta^{(1)}(\mathbf{z}_{ij}^T\boldsymbol{\alpha})g^{(2)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\\
\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\alpha}\partial\eta_{ij}} &= \mathbf{z}_{ij}\theta^{(1)}(\mathbf{z}_{ij}^T\boldsymbol{\alpha})g^{(2)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\\
\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \mathbf{x}_{ij}\mathbf{x}_{ij}^Tg^{(2)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),\\
\frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\beta}\partial\eta_{ij}} &= \mathbf{x}_{ij}g^{(2)}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\theta(\mathbf{z}_{ij}^T\boldsymbol{\alpha})),
\end{aligned}$$

and other second (mixed) derivatives are zero.

(b) The second derivatives of μ_{ijk} are

$$\frac{\partial^2\mu_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}=\begin{cases}\frac{1}{2(\psi_{ijk}-1)}\left[\frac{\partial^2a_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}-\frac{1}{2\sqrt{b_{ijk}}}\frac{\partial^2b_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}\right], & \psi_{ijk}\neq1,\\[2ex] \frac{\partial^2\mu_{ij}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}\mu_{ik}+2\frac{\partial\mu_{ij}}{\partial\boldsymbol{\zeta}}\frac{\partial\mu_{ik}}{\partial\boldsymbol{\zeta}^T}+\mu_{ij}\frac{\partial^2\mu_{ik}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}, & \psi_{ijk}=1,\end{cases}$$

$$\frac{\partial^2\mu_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\phi}^T}=\begin{cases}-\frac{1}{2(\psi_{ijk}-1)^2}\frac{\partial\psi_{ijk}}{\partial\boldsymbol{\phi}^T}\left[\frac{\partial a_{ijk}}{\partial\boldsymbol{\zeta}}-\frac{1}{2\sqrt{b_{ijk}}}\frac{\partial b_{ijk}}{\partial\boldsymbol{\zeta}}\right]\\ -\frac{1}{2(\psi_{ijk}-1)\sqrt{b_{ijk}}}\left[\frac{\partial^2a_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\phi}^T}+\frac{1}{4b_{ijk}}\cdot\frac{\partial b_{ijk}}{\partial\boldsymbol{\phi}^T}\cdot\frac{\partial b_{ijk}}{\partial\boldsymbol{\zeta}}-\frac{1}{2}\cdot\frac{\partial^2b_{ijk}}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\phi}^T}\right], & \psi_{ijk}\neq1,\\[2ex] \mathbf{0}, & \psi_{ijk}=1,\end{cases}$$

and

$$\frac{\partial^2 \mu_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} = \begin{cases} -\frac{1}{(\psi_{ijk}-1)^2} \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}^T} \left[-\mu_{ijk} \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}} + \frac{1}{2} \left(\frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}} - \frac{1}{2\sqrt{b_{ijk}}} \frac{\partial b_{ijk}}{\partial \boldsymbol{\phi}} \right) \right] \\ + \frac{1}{(\psi_{ijk}-1)} \left[-\frac{\partial \mu_{ijk}}{\partial \boldsymbol{\phi}^T} \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}} - \mu_{ijk} \frac{\partial^2 \psi_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \right. \\ \left. + \frac{1}{2} \left(\frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} + \frac{1}{4 b_{ijk} \sqrt{b_{ijk}}} \frac{\partial b_{ijk}}{\partial \boldsymbol{\phi}} \frac{\partial b_{ijk}}{\partial \boldsymbol{\phi}^T} - \frac{1}{2\sqrt{b_{ijk}}} \frac{\partial^2 b_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} \right) \right], & \psi_{ijk} \neq 1, \\ \mathbf{0}, & \psi_{ijk} = 1, \end{cases}$$

where

$$\begin{aligned}
\frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} &= -(1-\psi_{ijk}) \left(\frac{\partial^2 \mu_{ij}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} + \frac{\partial^2 \mu_{ik}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} \right), \\
\frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\phi}^T} &= \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}} \left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}^T} + \frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}^T} \right), \\
\frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} &= \frac{\partial^2 \psi_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} (\mu_{ij} + \mu_{ik}), \\
\frac{\partial^2 b_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} &= 2 \frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}} \frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}^T} + 2 a_{ijk} \frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} \\
&\quad -4\psi_{ijk}(1-\psi_{ijk}) \left(\frac{\partial^2 \mu_{ij}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} \mu_{ik} + 2 \frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}} \frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}^T} + \mu_{ij} \frac{\partial \mu_{ik}^T}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\zeta}^T} \right), \\
\frac{\partial^2 b_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\phi}^T} &= 2 \frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}^T} \frac{\partial a_{ijk}}{\partial \boldsymbol{\zeta}} + 2 a_{ijk} \frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\zeta} \partial \boldsymbol{\phi}^T} \\
&\quad -4(1-2\psi_{ijk}) \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}^T} \left(\frac{\partial \mu_{ij}}{\partial \boldsymbol{\zeta}} \mu_{ik} + \mu_{ij} \frac{\partial \mu_{ik}}{\partial \boldsymbol{\zeta}} \right), \\
\frac{\partial^2 b_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} &= 2 \frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}^T} \cdot \frac{\partial a_{ijk}}{\partial \boldsymbol{\phi}} + 2 a_{ijk} \frac{\partial^2 a_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T} - 8 \mu_{ij} \mu_{ik} \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}} \frac{\partial \psi_{ijk}}{\partial \boldsymbol{\phi}^T} \\
&\quad -4(2\psi_{ijk}-1) \mu_{ij} \mu_{ik} \frac{\partial^2 \psi_{ijk}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}.
\end{aligned}$$

References

Carey, V., Zeger, S. L. and Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517-526.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Connolly, M. and Liang, K. Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75**, 501-506.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

Cupples, L. A, Yang, Q., Demissie, S., Copenhafer, D. and Levy, D., (2003). Description of the Framingham Heart Study Data for Genetic Analysis Workshop 13. *BMC Genet* [Suppl] 4:2.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141-151.

Fitzmaurice, G. M. and Lipsitz, S. R. (1995). A model for binary time series data with serial odds ratio patterns. *Appl. Statist.* **44**, 51-61.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *J. Roy. Statist. Soc. Ser. B* **57**, 691-704.

Kraft, P., Bauman, L., Yuan, J. Y. and Horvath, S. (2003). Multivariate variance-components analysis of longitudinal blood pressure measurements from the Framingham Heart Study. *BMC Genetics* **4** (Suppl 1):S55.

Lang, J. and Agresti, A. A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**, 625-632.

Liang, H., Wang, S., Robins, J. M. and Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.* **99**, 357-367.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 3-40.

Lin, X. H. and Carroll, R. J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96**, 1045-1056.

Lin, X. H. and Carroll, R. J. (2001b). Semiparametric regression for clustered data. *Biometrika* **88**, 1179-1185.

Lin, X. H. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. Roy. Statist. Soc. Ser. B* **68**, 69-88.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221-239.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using an n-way Plackett distribution. *J. Amer. Statist. Assoc.* **89**, 633-644.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43-52.

Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered Data. *J. Amer. Statist. Assoc.* **100**, 147-157.

Yi, G. Y. and Cook, R. J. (2002a). Marginal methods for incomplete longitudinal data arising in clusters. *J. Amer. Statist. Assoc.* **97**, 1071-1080.

Yi, G. Y. and Cook, R. J. (2002b). Second order estimating equations for clustered longitudinal binary data with missing observations. In *Recent Advances in Statistical Methods* (Edited by Y. P. Chaubey), 352-366, World Scientific Publishing Company Inc., London.

Yi, G. Y., He, W. and Liang, H. (2009). Analysis of correlated binary data under partially linear single-index logistic models. *J. Multivariate Anal.* **100**, 278-290.

Yi, G. Y., He, W. and Liang, H. (2010). Semiparametric marginal and association regression methods for clustered binary data. To appear in *Ann. Inst. Statist. Math.*.

Yi, G. Y. and Thompson, M. E. (2005). Marginal and association regression models for longitudinal binary data with drop-outs: A likelihood-based approach. *Canad. J. Statist.* **33**, 3-20.

Yi, G. Y., Zeng, L. and Cook, R. J. (2010). A robust pairwise likelihood method for incomplete longitudinal binary data arising in clusters. To appear in *Canad. J. Statist.*.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 321-327.

Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7.

E-mail: whe@stats.uwo.ca

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

E-mail: yyi@uwaterloo.ca

(Received October 2009; accepted March 2010)