International Journal of Statistics and Probability; Vol. 4, No. 3; 2015
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Semiparametric Marginal Models for Binary Longitudinal Data

Salehin K. Chowdhury[1] & Sanjoy K. Sinha[1]

[1] School of Mathematics and Statistics, Carleton University, Canada

Correspondence: Salehin K. Chowdhury, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada. E-mail: schowdh2@math.carleton.ca

Received: May 7, 2015 Accepted: June 1, 2015 Online Published: July 10, 2015

doi:10.5539/ijsp.v4n3p107 URL: http://dx.doi.org/10.5539/ijsp.v4n3p107

This research is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Abstract

In this paper, we propose and explore a semiparametric approach to analyzing longitudinal binary data often observed in clinical studies. We applied second-order GEE approach to analyze longitudinal binary responses based on a partially linear single-index model. We use a local polynomial smoothing technique to estimate the single-index. We study the empirical properties of the proposed estimators using simulations. The empirical results demonstrate that if the true underlying model is partially linear, then our proposed method generally provides unbiased and efficient estimators. The proposed method is also applied to some real data sets obtained from longitudinal studies.

Keywords: Generalized estimating equation, Single-index model, Local polynomial smoothing, Longitudinal data, Marginal model, Single-index model.

1. Introduction

The commonly used linear regression paradigm models a dependent variable $\mathbf{Y}$ as a linear function of a vector-valued independent variable $\mathbf{x}$ in the form $E[\mathbf{Y}|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$, for a vector of unknown parameters $\boldsymbol{\beta}$. Generalized linear models (GLMs) provide a flexible extension of linear models by assuming that the dependent variable Y is of the form $E[\mathbf{Y}|\mathbf{x}] = \varphi(\mathbf{x}'\boldsymbol{\beta})$, for some known inverse link function or transfer function φ, which includes the commonly used logistic regression with $\varphi(z) = 1/(1 + e^{-z})$.

The case when both φ and $\boldsymbol{\beta}$ are unknown is referred to as single-index models (SIM). These models involve challenging problems of jointly estimating φ and $\boldsymbol{\beta}$, where φ may come from a large non-parametric family such as the family of all monotonic functions.

In practice, the relationship between the response and covariates may be very complex and linear terms may not be adequate to feature that relationship. Under these circumstances, it might be preferable to include a linear term $\mathbf{x}'\boldsymbol{\beta}$ and a nonlinear term $\varphi(\mathbf{u}'\boldsymbol{\xi})$ in a semiparametric regression, where φ is a smooth but unknown function. These models are a natural combination and generalization of simpler models already in the literature, namely single-index models and partially linear models. Carroll et al. (1997) called it generalized partially linear single-index models (GPLSIM). Li (1991) noted that if φ is monotone, then $\boldsymbol{\xi}$ takes on the same general meaning as "effect" parameters as would occur in ordinary linear models. Furthermore, it is a widely applied method to include a nonparametric function into the model for covariates $\mathbf{u}$ that have large dimension and are of little interest (e.g., confounders). This allows us to make inference on the effects of $\mathbf{x}$ while making minimal assumptions on $\mathbf{u}$. Marginal semiparametric models and their various extensions have become increasingly popular (see, for example, Carroll et al., 1997; Fan and Li, 2004; Wang et al., 2005; and Yi et al., 2009; among others). Carroll et al. (1997) propose estimates of the unknown parameters $(\boldsymbol{\beta}, \xi)$ and unknown function φ using quasi-likelihood and local polynomial regression and obtain their asymptotic distributions. However, they did not consider the joint estimation of the marginal mean parameters and association parameters.

In many applications, simply working on the marginal mean responses could be very restrictive. Estimation of the association parameters may be the central theme of the study. For example, in familial studies of inherited traits and developmental toxicology studies of laboratory animals, subjects in a family or cluster share common genetic traits or are subject to common environmental factors, and hence it is of prime scientific interest to study

the association between responses.

In this paper, we propose a method based on Prentice's (1988) second-order GEE approach to analyze longitudinal binary responses under partially linear single-index models. The computing algorithm for solving usual estimating equations based on the Newton-Raphson method cannot be directly employed due to the inclusion of a nonlinear function whose form is unknown. To circumvent this problem, following Carroll et al. (1997), we use a local polynomial smoothing technique to perform estimation of the single-index. We study the performance of the proposed method using simulations.

The paper is organized as follows. Section 2 introduces partially linear single-index models for analyzing binary longitudinal data. Section 3 describes the proposed second-order GEE approach for fitting single-index models. Section 4 discusses asymptotic properties of the proposed semiparametric GEE estimators. Section 5 presents results from a simulation study, which was carried out to investigate the empirical properties of the proposed estimators. Section 6 presents applications of the proposed method using two real data sets obtained from longitudinal studies. Section 7 gives the conclusions of the paper.question.

2. Single-Index Model for Binary Response

Suppose in a longitudinal study, there are N subjects and each subject is observed at a fixed set of T time-points. Let Y_{it} denote a binary response variable from subject i observed at time t for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. For subject i, we have a $T \times 1$ response vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$. Let y_{it} and $\mathbf{y}_i$ denote the realizations of the responses Y_{it} and $\mathbf{Y}_i$, respectively. Also, let $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})'$ and $\mathbf{u}_{it} = (u_{it1}, \ldots, u_{itq})'$ denote $p \times 1$ and $q \times 1$ vectors of covariates, respectively, from subject i at time t, which may be time-dependent or fixed across the observation times. Define $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$ and $\mathbf{u}_i = (\mathbf{u}'_{i1}, \ldots, \mathbf{u}'_{iT})'$ as the vectors of covariates for subject i.

Assume that the marginal distribution of Y_{it} is Bernoulli:

$$Y_{it} \sim \text{Bernoulli}(p_{it}),\ i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{1}$$

with the probability of success,

$$p_{it} = E(Y_{it}|\mathbf{x}_i, \mathbf{u}_i) = P(Y_{it} = 1|\mathbf{x}_i, \mathbf{u}_i). \tag{2}$$

Denote $\mathbf{p}_i = (p_{i1}, \ldots, p_{iT})'$. Assuming that the mean of response variable Y_{it} depends only on the covariate vector for subject i at time t, i.e., $p_{it} = E(Y_{it}|\mathbf{x}_i, \mathbf{u}_i) = E(Y_{it}|\mathbf{x}_{it}, \mathbf{u}_{it})$ (Pepe and Anderson, 1994), we consider modelling the mean response by a partially linear single-index logistic regression model in the form

$$\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) = \mathbf{x}'_{it}\boldsymbol{\beta} + \varphi(\mathbf{u}'_{it}\boldsymbol{\xi}), \quad \text{with} \quad \|\boldsymbol{\xi}\| = 1, \tag{3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are $p \times 1$ and $q \times 1$ unknown parameter vectors, respectively, and φ is an unknown smooth function. The restriction $\|\boldsymbol{\xi}\| = 1$ ensures identifiability of $\boldsymbol{\xi}$ (Carroll et al., 1997). Model (3) generalizes the usual logistic regression in the sense that a nonlinear term, $\varphi(\mathbf{u}'_{it}\boldsymbol{\xi})$, is included in the model. If $\varphi(.)$ is specified as the identity function, then model (3) becomes an ordinary logistic regression model with a known link function.

The marginal variance of the response variable Y_{it} is specified as a function of the marginal mean as

$$v_{it} = \text{var}(Y_{it}|\mathbf{x}_{it}) = p_{it}(1 - p_{it}). \tag{4}$$

We assume that Y_{it} and $Y_{i't'}$ are uncorrelated when $i \neq i'$. Let

$$\text{corr}(Y_{it}, Y_{it'}) = \alpha_{tt'} \tag{5}$$

represent the correlation between the responses Y_{it} and $Y_{it'}$ for the given vector of covariates $\mathbf{x}_i$. Denote $\boldsymbol{\alpha} = (\alpha_{12}, \ldots, \alpha_{1T}, \alpha_{23}, \ldots, \alpha_{T-1,T})'$ as the vector of correlation parameters.

3. Methods of Estimation

3.1 GEEs When φ is Identity

If one naively assumes that the single-index φ is identity, then the regression and association parameters may be estimated by using the ordinary GEEs (Liang and Zeger, 1986). In this case, estimates of the regression parameters $(\boldsymbol{\beta}, \boldsymbol{\xi})$ may be obtained by solving the estimating equations

$$\mathbf{U}_{\beta,\xi}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{D}'_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{p}_i) = \mathbf{0}, \tag{6}$$

where $\mathbf{D}_i = (\partial \mathbf{p}_i/\partial \boldsymbol{\beta}, \partial \mathbf{p}_i/\partial \boldsymbol{\xi})'$, $\mathbf{V}_i = \mathbf{B}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{B}_i^{1/2}$ is the marginal covariance matrix of $\mathbf{Y}_i$ with $\mathbf{B}_i = \text{diag}\{p_{i1}(1 - p_{i1}), p_{i2}(1 - p_{i2}) \ldots, p_{iT}(1 - p_{iT})\}$ and $\mathbf{R}(\boldsymbol{\alpha})$ is a correlation matrix for $\mathbf{Y}_i$ depending on the vector $\boldsymbol{\alpha}$ of correlation parameters. The above equations can be solved numerically for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\xi}}$ using an iterative method.

Liang and Zeger (1986) considered estimating the association parameters $\boldsymbol{\alpha}$ by the method of moments, which uses the Pearson residuals

$$\hat{r}_{it} = \frac{(y_{it} - \hat{p}_{it})}{(1 - \hat{p}_{it})^{1/2}}. \tag{7}$$

The moment estimators of $\alpha_{tt'}$ may be obtained as

$$\hat{\alpha}_{tt'} = \sum_{i=1}^{N} \hat{r}_{it}\hat{r}_{it'}/(N - (p + q)), \tag{8}$$

where p and q are the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$, respectively.

Prentice (1988) considered an extension of the GEE approach to allow joint estimation of the regression parameters ($\boldsymbol{\beta}$, $\boldsymbol{\xi}$) and the association parameters $\boldsymbol{\alpha}$. Specifically, a GEE estimator of the correlation parameter $\boldsymbol{\alpha}$ may be obtained from a second set of estimating equations by noting that the "sample correlation"

$$Z_{itu} = Z_{itu}(\beta) = \frac{(Y_{it} - p_{it})(Y_{iu} - p_{iu})}{(p_{it}q_{it}p_{iu}q_{iu})^{1/2}} \tag{9}$$

has mean ρ_{itu} and variance

$$w_{itu} = 1 + (1 - 2p_{it})(1 - 2p_{iu})(p_{it}q_{it}p_{iu}q_{iu})^{-1/2}\rho_{itu} - \rho_{itu}^2, \tag{10}$$

for $t < u < T$, $i = 1, \ldots, N$, and $t = 1, \ldots, T$. Let $\mathbf{Z}_i = (Z_{i12}, \ldots, Z_{i1T}, Z_{i23}, \ldots, Z_{i,T-1,T})'$ and $\boldsymbol{\rho}_i = (\rho_{i12}, \ldots, \rho_{i1T}, \rho_{i23}, \ldots, \rho_{i,T-1,T})'$.

Then following Prentice (1988), the second-order GEE (GEE2) estimators of $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ may be obtained by solving the estimating equations

$$\mathbf{U}_{\beta,\xi}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{p}_i) = \mathbf{0}, \tag{11}$$

$$\mathbf{U}_{\alpha}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{G}_i'\mathbf{W}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\rho}_i) = \mathbf{0}, \tag{12}$$

where $\mathbf{G}_i = \partial \boldsymbol{\rho}_i/\partial \boldsymbol{\alpha}$ and $\mathbf{W}_i = \text{diag}\{w_{i12}, \ldots, w_{i1T}, w_{i23}, \ldots, w_{i,T-1,T}\}$.

3.2 Semiparametric GEEs When φ is Unknown

An unknown smooth function φ leads to a partially linear single-index model, as defined in (3). For this, the estimating functions $\mathbf{U}_{\beta,\xi}$ and $\mathbf{U}_{\alpha}$ in (11) and (12) involve the unknown function φ. We propose to use a nonparametric approach to estimating φ locally prior to estimating the regression parameters ($\boldsymbol{\beta}$, $\boldsymbol{\xi}$) and association parameters $\boldsymbol{\alpha}$.

Assuming that the function $\varphi(c)$ has a second derivative, we may approximate $\varphi(c)$ by a locally linear function within the neighborhood of c_0 via the Taylor series expansion, $\varphi(c) \approx \varphi(c_0) + \varphi^{(1)}(c_0)(c - c_0)$, for a given point c_0, where $\varphi^{(1)}(c_0)$ is the first-order differentiation of $\varphi(c)$ with respect to c evaluated at c_0. Denote $\eta_0(c_0) = \varphi(c_0)$, $\eta_1(c_0) = \varphi^{(1)}(c_0)$, and $\boldsymbol{\eta}(c_0) = (\eta_0(c_0), \eta_1(c_0))'$.

We also denote $C_{it} = \mathbf{u}_{it}'\boldsymbol{\xi}$, $\boldsymbol{\phi}_{it}(c, \boldsymbol{\xi}) = (1, C_{it} - c)'$, $\Phi_i(c)$ is a $t \times 2$ matrix with the tth row $\boldsymbol{\phi}_{it}(c, \boldsymbol{\xi})$, and $\Delta_i = \text{diag}\{p_{it}^{(1)}, t = 1, \ldots, T\}$, where $p_{it}^{(1)}$ is the first-order derivative of the mean function p_{it} evaluated at $\mathbf{x}_{it}'\boldsymbol{\beta} + \varphi(\mathbf{u}_{it}'\boldsymbol{\xi})$. Let $K(c)$ be a kernel function (or a symmetric density function) with a compact support and h be a bandwidth. Denote $K_h(a) = K(a/h)/h$ and $\mathbf{K}_{ih}(c) = \text{diag}\{K_h(C_{it} - c), t = 1, \ldots, T\}$.

We describe below an algorithm for simultaneous estimation of the unknown function φ, mean parameters ($\boldsymbol{\beta}$, $\boldsymbol{\xi}$) and correlation parameters $\boldsymbol{\alpha}$. We adopt Prentice's (1988) second-order GEE approach to conduct simultaneous estimation of ($\boldsymbol{\beta}$, $\boldsymbol{\xi}$) and $\boldsymbol{\alpha}$.

Step 0. (Initialization step). Fit a parametric logistic model assuming $\varphi = 1$ to obtain initial values $(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0, \boldsymbol{\alpha}_0)$ and set $\hat{\boldsymbol{\xi}} = \boldsymbol{\xi}_0/\|\boldsymbol{\xi}_0\|$, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$.

Step 1. For a given point c in a selected grid, find $\hat{\varphi}(c, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}}) = \hat{\eta}_0(c)$ and $\hat{\eta}_1(c)$ by solving the equations

$$\sum_{i=1}^{N} \widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \widetilde{\mathbf{p}}_i\right) = \mathbf{0}, \tag{13}$$

with respect to $\boldsymbol{\eta}(c)$, where $\widetilde{\Phi}_i(c)$ and $\widetilde{\mathbf{K}}_{ih}(c)$ are $\Phi_i(c)$ and $\mathbf{K}_{ih}(c)$, respectively, with $\boldsymbol{\xi}$ replaced by $\hat{\boldsymbol{\xi}}$, $\widetilde{\mathbf{p}}_i = (\widetilde{p}_{i1}, \ldots, \widetilde{p}_{iT})$ with $\widetilde{p}_{it} = g\{\mathbf{x}_{it}'\hat{\boldsymbol{\beta}} + \boldsymbol{\phi}_{it}(c, \hat{\boldsymbol{\xi}})\boldsymbol{\eta}(c)\}$, $g(a) = \exp(a)/\{1 + \exp(a)\}$, $\widetilde{\Delta}_i$ is Δ_i with p_{it} replaced by $\widetilde{p}_{it}$, and $\widetilde{\mathbf{V}}_i$ is a "working independence" diagonal matrix with diagonal elements $\widetilde{V}_{it} = \widetilde{p}_{it}(1 - \widetilde{p}_{it})$.

Step 2. Given the estimate $\hat{\varphi}(c, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}}) = \hat{\eta}_0(c)$ and $\hat{\eta}_1(c)$ for points c in the selected grid, update $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}})$ by solving the following two sets of estimating equations for $(\boldsymbol{\beta}, \boldsymbol{\xi})$ and $\boldsymbol{\alpha}$:

$$\tilde{\mathbf{U}}_{\beta,\xi}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \hat{\varphi}) = \sum_{i=1}^{N} \hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\left\{\mathbf{Y}_i - \hat{\mathbf{p}}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})\right\} = \mathbf{0}, \tag{14}$$

$$\tilde{\mathbf{U}}_{\alpha}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \hat{\varphi}) = \sum_{i=1}^{N} \mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}(\hat{\mathbf{Z}}_i - \boldsymbol{\rho}_i) = \mathbf{0}, \tag{15}$$

where the mean vector $\hat{\mathbf{p}}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = (\hat{p}_{i1}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}), \ldots, \hat{p}_{iT}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}))'$ with its (i,t)th element $\hat{p}_{it}(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = g\left(\mathbf{x}_{it}'\boldsymbol{\beta} + \hat{\varphi}(\mathbf{u}_{it}'\boldsymbol{\xi}; \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})\right)$, $\hat{\mathbf{D}}_i$, $\hat{\mathbf{V}}_i$, $\hat{\mathbf{Z}}_i$ and $\hat{\mathbf{W}}_i$ are the same as $\mathbf{D}_i$, $\mathbf{V}_i$, $\mathbf{Z}_i$ and $\mathbf{W}_i$ in (11) and (12) respectively with $\mathbf{p}_i$ replaced by $\hat{\mathbf{p}}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})$. The vector $\boldsymbol{\rho}_i$ is the same as in Eq.(12).

Step 3. Repeat Steps 1 and 2 until the convergence of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}})$. The estimates at convergence are referred to as the semiparametric second-order GEE (SGEE2) estimates.

When implementing the aforementioned algorithm, it is often feasible to obtain a set of initial values by fitting a model with the ordinary GEE approach assuming that the single-index φ is the identity function. Our computational experience suggest that the algorithm is not severely but somewhat sensitive to the initial values, and the above choice of initial values ensures a convergence in estimation in most cases.

We conclude this section with a discussion on the bandwidth selection. As the bandwidth h affects both the bias and variance of an estimator, there is always a trade-off between these two inference criteria. Bias correction requires the choice of a relatively small bandwidth, whereas a smaller variance estimate needs a larger value of the bandwidth. In principle, the bandwidth selection is data driven, and traditional methods such as the cross-validation approach may be applied to select a proper bandwidth h based on available data. Carroll et al. (1997) noted that a sensible choice of the bandwidth h is generally difficult. Instead, they suggested an ad hoc bandwidth, given by $\hat{h}_{\text{opt}} \times N^{-2/15} = O(N^{-1/3})$, which satisfies $Nh^4 \to 0$ and $Nh^2/(\log N)^2 \to \infty$.

3.3 Computational Details for Semiparametric GEEs

To estimate the regression parameters $(\boldsymbol{\beta}, \boldsymbol{\xi})$ and association parameters $\boldsymbol{\alpha}$, in Step 1 of our algorithm, we estimate the single-index parameter φ by solving the estimating equations (13). For this, we can write

$$\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1} = \begin{bmatrix} K_h(C_{i1} - c) & \cdots & K_h(C_{iT} - c) \\ (C_{i1} - c)K_h(C_{i1} - c) & \cdots & (C_{iT} - c)K_h(C_{iT} - c) \end{bmatrix}, \tag{16}$$

and

$$\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \widetilde{\mathbf{p}}_i\right) = \begin{bmatrix} \sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - \tilde{p}_{it}) \\ \sum_{t=1}^{T}(C_{it} - c)K_h(C_{it} - c)(y_{it} - \tilde{p}_{it}) \end{bmatrix}, \tag{17}$$

where $\log\{\tilde{p}_{it}/(1 - \tilde{p}_{it})\} = \mathbf{x}_i'\boldsymbol{\beta} + \eta_0(c) + (C_{it} - c)\eta_1(c)$, and $K(.)$ is a kernel density function (e.g., standard normal density). We use a Newton-Raphson iterative algorithm for estimating φ. For this, we find

$$\frac{\partial(y_{it} - \tilde{p}_{it})}{\partial\eta_0(c)} = -\tilde{p}_{it}(1 - \tilde{p}_{it}), \tag{18}$$

$$\frac{\partial(y_{it} - \tilde{p}_{it})}{\partial\eta_1(c)} = -(C_{it} - c)\tilde{p}_{it}(1 - \tilde{p}_{it}), \tag{19}$$

and

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\eta}(c)}\left\{\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i-\widetilde{\mathbf{p}}_i\right)\right\}\\
&\approx \begin{bmatrix} \sum_{t=1}^{T} K_h(C_{it}-c)\frac{\partial(y_{it}-\tilde{p}_{it})}{\partial\eta_0} & \sum_{t=1}^{T}(C_{it}-c)K_h(C_{it}-c)\frac{\partial(y_{it}-\tilde{p}_{it})}{\partial\eta_0} \\ \sum_{t=1}^{T} K_h(C_{it}-c)\frac{\partial(y_{it}-\tilde{p}_{it})}{\partial\eta_1} & \sum_{t=1}^{T}(C_{it}-c)K_h(C_{it}-c)\frac{\partial(y_{it}-\tilde{p}_{it})}{\partial\eta_1} \end{bmatrix}\\
&= -\begin{bmatrix} \sum_{t=1}^{T} K_h(C_{it}-c)\tilde{p}_{it}(1-\tilde{p}_{it}) & \sum_{t=1}^{T}(C_{it}-c)K_h(C_{it}-c)\tilde{p}_{it}(1-\tilde{p}_{it}) \\ \sum_{t=1}^{T} K_h(C_{it}-c)\tilde{p}_{it}(1-\tilde{p}_{it}) & \sum_{t=1}^{T}(C_{it}-c)^2K_h(C_{it}-c)\tilde{p}_{it}(1-\tilde{p}_{it}) \end{bmatrix}\\
&= -\begin{bmatrix} K_h(C_{i1}-c) & \cdots & K_h(C_{iT}-c) \\ (C_{i1}-c)K_h(C_{i1}-c) & \cdots & (C_{iT}-c)K_h(C_{iT}-c) \end{bmatrix}\times\\
&\quad \begin{bmatrix} \tilde{p}_{i1}(1-\tilde{p}_{i1}) & (C_{i1}-c)\tilde{p}_{i1}(1-\tilde{p}_{i1}) \\ \vdots & \vdots \\ \tilde{p}_{iT}(1-\tilde{p}_{iT}) & (C_{iT}-c)\tilde{p}_{iT}(1-\tilde{p}_{iT}) \end{bmatrix}\\
&\equiv -\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\boldsymbol{H}, \qquad (20)
\end{aligned}$$

where the matrix $\boldsymbol{H}$ is defined by $\boldsymbol{H} = \partial\widetilde{\mathbf{p}}_i/\partial\boldsymbol{\eta}(c)$.

Given a set of current estimates $\hat{\boldsymbol{\eta}}_{(s)} = \{\hat{\eta}_{0(s)}(c), \hat{\eta}_{1(s)}(c)\}'$ of $\boldsymbol{\eta} = (\eta_0, \eta_1)'$, we can obtain updated estimates $\hat{\boldsymbol{\eta}}_{(s+1)} = \{\hat{\eta}_{0(s+1)}(c), \hat{\eta}_{1(s+1)}(c)\}'$ from the iterative equations

$$\hat{\boldsymbol{\eta}}_{(s+1)} = \hat{\boldsymbol{\eta}}_{(s)} + \left(\sum_{i=1}^{N}\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\boldsymbol{H}\right)^{-1}\sum_{i=1}^{N}\widetilde{\Phi}_i'(c)\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i-\widetilde{\mathbf{p}}_i\right), \qquad (21)$$

for $s = 0, 1, 2, \ldots$, where the second term on the right side is evaluated at the current estimates $\hat{\boldsymbol{\eta}}_{(s)} = \{\hat{\eta}_{0(s)}(c), \hat{\eta}_{1(s)}(c)\}'$.

For a given estimate $\hat{\varphi}(c; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}}) = \hat{\eta}_0(c)$ and $\hat{\eta}_1(c)$, the success probability p_{it} may be estimated from

$$\text{logit}(\hat{p}_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \hat{\eta}_0(c), \qquad (22)$$

and also we can find

$$\frac{\partial\hat{p}_{it}}{\partial\xi_j} = u_{itj}\,\hat{\eta}_1(c)\hat{p}_{it}(1-\hat{p}_{it}). \qquad (23)$$

The iterative procedure for estimating $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ begins with some initial values $(\hat{\boldsymbol{\beta}}_{(0)}', \hat{\boldsymbol{\xi}}_{(0)}', \hat{\boldsymbol{\alpha}}_{(0)}')$ in Step 0 of our algorithm, and produces updated values $(\hat{\boldsymbol{\beta}}_{(s+1)}', \hat{\boldsymbol{\xi}}_{(s+1)}', \hat{\boldsymbol{\alpha}}_{(s+1)}')$ from the iterative equations

$$\left(\hat{\boldsymbol{\beta}}_{(s+1)}', \hat{\boldsymbol{\xi}}_{(s+1)}'\right)' = \left(\hat{\boldsymbol{\beta}}_{(s)}', \hat{\boldsymbol{\xi}}_{(s)}'\right)' + \left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\left\{\mathbf{Y}_i-\hat{\mathbf{p}}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\alpha})\right\}, \qquad (24)$$

and

$$\hat{\boldsymbol{\alpha}}_{(s+1)} = \hat{\boldsymbol{\alpha}}_{(s)} + \left(\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i\right)^{-1}\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right), \qquad (25)$$

for $s = 0, 1, 2, \ldots$, where the second terms on the right side are evaluated at the current estimates $(\hat{\boldsymbol{\beta}}_{(s)}, \hat{\boldsymbol{\xi}}_{(s)}, \hat{\boldsymbol{\alpha}}_{(s)})$

4. Asymptotic Properties

4.1 Asymptotics for $\hat{\varphi}$

The asymptotic distribution of the single-index estimator $\hat{\varphi}(c)$ may be established under the following assumptions (see Carroll et al., 1997, for details):

i) The density function of $\mathbf{u}_{it}$ has a continuous second derivative on its support.

ii) The density function of $\mathbf{u}_{it}'\boldsymbol{\xi}$ is positive and uniformly continuous for $\boldsymbol{\xi}$ in a neighborhood of its true value.

iii) The second-order derivative $\varphi^{(2)}(c)$ is continuous on its support.

iv) The random vector $\mathbf{x}_{it}$ is assumed to have a bounded support with $E(\mathbf{x}_{it}'\mathbf{x}_{it}) > 0$.

v) $K(.)$ is a symmetric probability density function with bounded support.

From Eq. (13) we can write,

$$\begin{aligned}
&\sum_{i=1}^{N} \mathbf{1}'\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \widetilde{\mathbf{p}}_i\right) = 0 \\
\Rightarrow &\sum_{i=1}^{N} \mathbf{1}'\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\mathbf{Y}_i - \mathbf{p}_i\right) - \sum_{i=1}^{N} \mathbf{1}'\widetilde{\Delta}_i\widetilde{\mathbf{K}}_{ih}(c)\widetilde{\mathbf{V}}_i^{-1}\left(\widetilde{\mathbf{p}}_i - \mathbf{p}_i\right) = 0 \\
\Rightarrow &\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - p_{it}) = \sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)(\tilde{p}_{it} - p_{it})
\end{aligned} \tag{26}$$

Denote the right side of (26) by $\sum_{i=1}^{N} \Phi_i^*(\hat{\eta}_0(c))$, where

$$\Phi_i^*(\hat{\eta}_0(c)) = \sum_{t=1}^{T} K_h(C_{it} - c)(\tilde{p}_{it} - p_{it}).$$

Then by a Taylor series expansion, we get

$$\begin{aligned}
\sum_{i=1}^{N} \Phi_i^*(\hat{\eta}_0(c)) &\approx \sum_{i=1}^{N} \Phi_i^*(\eta_0(c)) + \sum_{i=1}^{N} \frac{\partial}{\partial\hat{\eta}_0(c)}\Phi_i^*(\hat{\eta}_0(c))\Big|_{\hat{\eta}_0(c)=\eta_0(c)} \{\hat{\eta}_0(c) - \eta_0(c)\} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)p_{it}(1 - p_{it})\{\hat{\eta}_0(c) - \eta_0(c)\}.
\end{aligned} \tag{27}$$

Thus from (26) and (27), we get

$$\begin{aligned}
\hat{\varphi}(c) - \varphi(c) &= \hat{\eta}_0(c) - \eta_0(c) \\
&\cong \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - p_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)p_{it}(1 - p_{it})}.
\end{aligned} \tag{28}$$

Let the marginal density of $\mathbf{u}'\boldsymbol{\xi}$ be denoted by $f(.)$. Applying the asymptotic properties of the kernel estimators and the Taylor series expansion, following Carroll et al. (1997) and Yi et al. (2009), we find the asymptotic expansion

$$\begin{aligned}
\hat{\varphi}(c) - \varphi(c) = \frac{\int a^2K(a)da}{2}\varphi^{(2)}(c)h^2 + \frac{1}{Nf(c)}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(C_{it} - c)\Omega_{it} \\
+o_p\{h^2 + (Nh)^{-1/2}\},
\end{aligned} \tag{29}$$

where Ω_{it} is the first element of the vector $(Y_{it} - p_{it})\boldsymbol{\Omega}^{-1}(c)(1, \mathbf{x}_{it}')'$ and

$$\boldsymbol{\Omega}(c) = E\left[p_{it}(1 - p_{it})\begin{pmatrix} 1 & \mathbf{x}_{it}' \\ \mathbf{x}_{it} & \mathbf{x}_{it}\mathbf{x}_{it}' \end{pmatrix}\middle| C_{it} = c\right]. \tag{30}$$

As a consequence, the asymptotic distribution of the single-index estimator $\hat{\varphi}(c)$ is obtained as

$$(Nh)^{1/2}\left\{\hat{\varphi}(c) - \varphi(c) - \frac{\int a^2K(a)dt}{2}\varphi^{(2)}(c)h^2\right\} \to_d \mathrm{N}\left(0, \frac{d(c)}{f(c)}\int K^2(a)da\right), \tag{31}$$

where $d(c)$ is the first diagonal element of the matrix $\boldsymbol{\Omega}^{-1}(c)$.

4.2 Asymptotics for $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\alpha}})$

Applying the first-order Taylor series approximation, from Eqs.(14)–(15) the vector of estimators $N^{1/2}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})', (\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})', (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})'\right]'$ may be approximated by

$$\left[\begin{array}{cc} -N^{-1}\frac{\partial \tilde{\mathbf{U}}_{\beta,\xi}(\beta,\xi,\alpha,\hat{\varphi})}{\partial(\beta,\xi)'} & -N^{-1}\frac{\partial \tilde{\mathbf{U}}_{\beta,\xi}(\beta,\xi,\alpha,\hat{\varphi})}{\partial\alpha'} \\ -N^{-1}\frac{\partial \tilde{\mathbf{U}}_{\alpha}(\beta,\xi,\alpha,\hat{\varphi})}{\partial(\beta,\xi)'} & -N^{-1}\frac{\partial \tilde{\mathbf{U}}_{\alpha}(\beta,\xi,\alpha,\hat{\varphi})}{\partial\alpha'} \end{array}\right]^{-1}\left[\begin{array}{c} N^{-1/2}\tilde{\mathbf{U}}_{\beta,\xi}(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi}) \\ N^{-1/2}\tilde{\mathbf{U}}_{\alpha}(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi}) \end{array}\right] \equiv \boldsymbol{J}^{-1}\boldsymbol{L}. \tag{32}$$

From this, we have

$$\mathrm{Var}\left(\boldsymbol{J}^{-1}\boldsymbol{L}\right) = \boldsymbol{J}^{-1}\mathrm{Var}(\boldsymbol{L})(\boldsymbol{J}^{-1})', \tag{33}$$

with

$$\begin{aligned} \mathrm{Var}(\boldsymbol{L}) &= N^{-1}\left[\begin{array}{cc} \mathrm{Var}(\tilde{\mathbf{U}}_{\beta,\xi}) & \mathrm{Cov}(\tilde{\mathbf{U}}_{\beta,\xi},\tilde{\mathbf{U}}_{\alpha}) \\ \mathrm{Cov}(\tilde{\mathbf{U}}_{\beta,\xi},\tilde{\mathbf{U}}_{\alpha}) & \mathrm{Var}(\tilde{\mathbf{U}}_{\alpha}) \end{array}\right] \\ &= N^{-1}\left[\begin{array}{cc} \sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\mathrm{Var}(\mathbf{Y}_i)\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i & \sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\mathrm{Cov}(\mathbf{Y}_i,\hat{\mathbf{Z}}_i)\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i \\ \sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathrm{Cov}(\mathbf{Y}_i,\hat{\mathbf{Z}}_i)\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i & \sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathrm{Var}(\hat{\mathbf{Z}}_i)\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i \end{array}\right] \\ &\equiv N^{-1}\left[\begin{array}{cc} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{array}\right], \end{aligned} \tag{34}$$

where the variance and covariance terms may be approximated by

$$\begin{aligned} \mathrm{Var}(\mathbf{Y}_i) &\approx (\mathbf{Y}_i-\hat{\mathbf{p}}_i)(\mathbf{Y}_i-\hat{\mathbf{p}}_i)' \\ \mathrm{Cov}(\mathbf{Y}_i,\hat{\mathbf{Z}}_i) &\approx (\mathbf{Y}_i-\hat{\mathbf{p}}_i)\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)' \\ \mathrm{Var}(\hat{\mathbf{Z}}_i) &\approx \left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)\left(\hat{\mathbf{Z}}_i-\hat{\boldsymbol{\rho}}_i\right)'. \end{aligned} \tag{35}$$

Moreover, if $\hat{\varphi}$ is consistent, then the linear functions $\boldsymbol{L}$ generally have an asymptotic multivariate normal distribution with mean vector zero and covariance matrix

$$\boldsymbol{\Sigma}_L = \lim_{N\to\infty} N^{-1}\left(\begin{array}{cc} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{array}\right). \tag{36}$$

Thus the joint asymptotic distribution of $N^{1/2}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})', (\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})', (\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})'\right]'$ is multivariate normal with mean vector zero and covariance matrix

$$\boldsymbol{\Sigma} = \lim_{N\to\infty}\boldsymbol{J}^{-1}\boldsymbol{\Sigma}_L\lim_{N\to\infty}(\boldsymbol{J}^{-1})'. \tag{37}$$

Also as $N\to\infty$,

$$\begin{aligned} -N^{-1}\frac{\partial\tilde{\mathbf{U}}_{\beta,\xi}(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi})}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} &= -N^{-1}\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\frac{\partial(\mathbf{Y}_i-\hat{\mathbf{p}}_i)}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} - N^{-1}\sum_{i=1}^{N}\frac{\partial\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'}(\mathbf{Y}_i-\hat{\mathbf{p}}_i) \\ &= N^{-1}\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i + o_p(1), \end{aligned} \tag{38}$$

since $\partial(\mathbf{Y}_i-\hat{\mathbf{p}}_i)/\partial(\boldsymbol{\beta},\boldsymbol{\xi})' = (-1)\hat{\mathbf{D}}_i$ and $\partial\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}/\partial(\boldsymbol{\beta},\boldsymbol{\xi})'$ are fixed matrices that do not involve $\mathbf{Y}_i$. Similarly, as $\partial\hat{\mathbf{V}}_i^{-1}/\partial\boldsymbol{\alpha}'$, $\partial\mathbf{E}_i'\hat{\mathbf{W}}_i^{-1}/\partial\boldsymbol{\alpha}'$ and $\partial\mathbf{E}_i'\hat{\mathbf{W}}_i^{-1}/\partial(\boldsymbol{\beta},\boldsymbol{\xi})'$ are all free of $\mathbf{Y}_i$, we can write as $N\to\infty$,

$$-N^{-1}\frac{\partial\tilde{\mathbf{U}}_{\beta,\xi}(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi})}{\partial\boldsymbol{\alpha}'} = -N^{-1}\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\frac{\partial\hat{\mathbf{V}}_i^{-1}}{\partial\boldsymbol{\alpha}'}(\mathbf{Y}_i-\hat{\mathbf{p}}_i) = o_p(1), \tag{39}$$

$$-N^{-1}\frac{\partial\tilde{\mathbf{U}}_\alpha(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi})}{\partial\boldsymbol{\alpha}'} = -N^{-1}\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\frac{\partial\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)}{\partial\boldsymbol{\alpha}'} - N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}}{\partial\boldsymbol{\alpha}'}\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)$$
$$= N^{-1}\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i + o_p(1), \tag{40}$$

and

$$-N^{-1}\frac{\partial\tilde{\mathbf{U}}_\alpha(\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\alpha},\hat{\varphi})}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} = -N^{-1}\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\frac{\partial\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} - N^{-1}\sum_{i=1}^{N}\frac{\partial\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'}\left(\hat{\mathbf{Z}}_i-\boldsymbol{\rho}_i\right)$$
$$= -N^{-1}\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\frac{\partial\hat{\mathbf{Z}}_i}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} + o_p(1). \tag{41}$$

We can show that

$$\frac{\partial\hat{Z}_{itu}}{\partial\beta_j} = -\left\{\frac{\partial\hat{p}_{it}}{\partial\beta_j}(Y_{iu}-\hat{p}_{iu}) + \frac{\partial\hat{p}_{iu}}{\partial\beta_j}(Y_{it}-\hat{p}_{it}) + \frac{1}{2}(Y_{it}-\hat{p}_{it})(Y_{iu}-\hat{p}_{iu})\right.$$
$$\left.\times\left[(1-2\hat{p}_{it})\hat{p}_{it}^{-1}\hat{q}_{it}^{-1}\frac{\partial\hat{p}_{it}}{\partial\beta_j} + (1-2\hat{p}_{iu})\hat{p}_{iu}^{-1}\hat{q}_{iu}^{-1}\frac{\partial\hat{p}_{iu}}{\partial\beta_j}\right]\right\}(\hat{p}_{it}\hat{q}_{it}\hat{p}_{iu}\hat{q}_{iu})^{-1/2}. \tag{42}$$

To find the derivative $\partial\hat{Z}_{itu}/\partial\xi_j$ in (42), we can replace $\partial\hat{p}_{it}/\partial\beta_j$ and $\partial\hat{p}_{iu}/\partial\beta_j$ by $\partial\hat{p}_{it}/\partial\xi_j$ and $\partial\hat{p}_{iu}/\partial\xi_j$, respectively. Hence as $N\to\infty$, we can write

$$\boldsymbol{J} = N^{-1}\begin{bmatrix}\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i & \mathbf{0}\\ -\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\frac{\partial\hat{\mathbf{Z}}_i}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'} & \sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i\end{bmatrix} \tag{43}$$

and

$$\boldsymbol{J}^{-1} = N\begin{bmatrix}\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} & \mathbf{0}\\ \boldsymbol{B} & \left(\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i\right)^{-1}\end{bmatrix}, \tag{44}$$

where

$$\boldsymbol{B} = \left(\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\mathbf{G}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{G}_i'\hat{\mathbf{W}}_i^{-1}\frac{\partial\hat{\mathbf{Z}}_i}{\partial(\boldsymbol{\beta},\boldsymbol{\xi})'}\right)\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}. \tag{45}$$

5. Simulation Study

To study the relative performance of the proposed SGEE2 method as compared to the ordinary GEE2 method, we ran two sets of simulations. In the first set, the estimators were studied for the case when the true model is a linear single-index model. In the second set, we considered the true model as a partially linear single-index model.

5.1 Response Model

We generated the data by considering a two-group design configuration with a binary response measured on four occasions. The marginal mean response $E(Y_{it}|\mathbf{x}_{it},\mathbf{u}_{it},\beta,\boldsymbol{\xi},\alpha)$ was defined by

$$\text{logit}\{E(Y_{it}|\mathbf{x}_{it},\mathbf{u}_{it},\beta,\boldsymbol{\xi},\alpha)\} = \beta x_i + \varphi(\xi_1u_{1it}+\xi_2u_{2it}+\xi_3u_{3it}),\quad t=1,\ldots,4, \tag{46}$$

where x_i is a dichotomous covariate indicating the group membership for the ith individual ($i=1,\ldots,N$) observed over a fixed set of $T=4$ time-points. We considered $P(X_i=1)=0.5$ throughout the simulations.

The data were generated from two different models. In the first model, covariates u_{ij}'s were generated from the uniform distribution U(0, 1), and parameters were fixed at $\beta=1$ and $\boldsymbol{\xi}=(\xi_1,\xi_2,\xi_3)'=(1/\sqrt{3},-1/\sqrt{3},1/\sqrt{3})'$. In the second model, covariates $\mathbf{u}_i$'s were chosen as the indicators of time-points, with $u_{jit}=1$ if $t=j\,(j=1,2,3)$ and $u_{jit}=0$ otherwise. The parameters were fixed at $\beta=1$ and $\boldsymbol{\xi}=(\xi_1,\xi_2,\xi_3)'=(1/\sqrt{3},1/\sqrt{3},1/\sqrt{3})'$.

Table 1. Comparison of Prentice's (1988) ordinary second-order GEE (GEE2) with the proposed semiparametric approach (SGEE2) when $\mathbf{u}_i$'s are continuous. [True parameter values: $\beta_1 = 1$, $\boldsymbol{\xi} = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})'$ and $\alpha = 0.3$.]

		Fitted Model			
		GEE2		SGEE2	
True Model	Parameters	Bias	MSE	Bias	MSE
Linear	β_1	–0.0056	0.0731	0.0063	0.0843
	ξ_1	0.0125	0.1081	0.0147	0.1231
	ξ_2	–0.0017	0.0921	–0.0846	0.0942
	ξ_3	0.0058	0.0924	0.0734	0.0860
	α	–0.0020	0.0027	–0.0037	0.0028
Partially linear	β_1	–0.0273	0.0592	0.0384	0.0934
	ξ_1	–0.2678	0.1591	–0.0030	0.0254
	ξ_2	0.0820	0.1147	–0.0679	0.0207
	ξ_3	–0.2590	0.1639	–0.0032	0.0252
	α	–0.0056	0.0029	0.0034	0.0022

To assess the performance of our proposed semiparametric (SGEE2) approach, we compare it with the ordinary second-order GEE (GEE2) approach of Prentice (1988) under two situations of data structures. First, we consider a scenario where a standard GEE model well fits the data with a linear single-index model. As a linear function is a special case of a nonlinear function, we expect that the proposed estimators would still be consistent, but there may be a possible efficiency loss incurred. In order to generate the data, φ is specified as the identity function. Second, we consider a scenario when our proposed method (SGEE2) well fits the data with a partially linear single-index model where we chose $\varphi(a) = \sin(\pi(1-a))$.

Throughout the simulations, data were generated under exchangeable correlation structures among the responses and we chose $\text{corr}(Y_{it}, Y_{it'}) = \alpha = 0.3$ for all $(t \neq t' = 1, \ldots, T)$. Each simulation run was based on 1000 replications of data sets, with each data set containing $N = 100$ subjects and $T = 4$ observations per subject.

5.2 Diagnostic Methods

The two methods were compared based on empirical biases and mean squared errors (MSEs) of the estimates. The bias of an estimator $\hat{\theta}$ of θ is estimated by

$$\text{bias}(\hat{\theta}) \approx \sum_{s=1}^{S} \frac{(\hat{\theta}_s - \theta)}{S}, \tag{47}$$

where $\hat{\theta}_s$ is the estimate of θ obtained from the sth simulated data set and S is the simulation size.The mean squared error (MSE) of $\hat{\theta}$ is estimated by

$$\text{MSE}(\hat{\theta}) \approx \sum_{s=1}^{S} \frac{(\hat{\theta}_s - \theta)^2}{S}. \tag{48}$$

5.2 Results

Table 1 presents the empirical biases and mean squared errors (MSEs) of the estimators of the regression parameters $(\beta, \xi_1, \xi_2, \xi_3)$ and the correlation parameter α for continuous covariates $\mathbf{u}_i$'s under both methods (GEE2 and SGEE2). Table 2 repeats these results for discrete covariates $\mathbf{u}_i$'s.

Both Table 1 and 2 show very similar results from the two methods when the true model is single-index linear. Although the finite sample biases from GEE2 tend to be smaller than those obtained from SGEE2, the biases from SGEE2 are still reasonably small. Also we notice that SGEE2 tends to produce larger MSE as compared to GEE2, but the differences do not seem considerable. In summary, if a nonparametric function is included to the mean

Table 2. Comparison of Prentice's (1988) ordinary second-order GEE (GEE2) with the proposed semiparametric approach (SGEE2) when $\mathbf{u}_i$'s are discrete. [True parameter values: $\beta_1 = 1, \boldsymbol{\xi} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})'$ and $\alpha = 0.3$.]

		Fitted Model			
		GEE2		SGEE2	
True Model	Parameters	Bias	MSE	Bias	MSE
Linear	β_1	–0.0525	0.0360	–0.0543	0.0432
	ξ_1	0.0219	0.0396	0.0436	0.0504
	ξ_2	0.0119	0.0534	–0.0382	0.0340
	ξ_3	0.0044	0.0438	–0.0326	0.0285
	α	–0.0027	0.0027	–0.0033	0.0026
Partially linear	β_1	–0.0321	0.0816	0.0107	0.0898
	ξ_1	0.0076	0.0535	–0.0020	0.0499
	ξ_2	–0.0175	0.0578	–0.0415	0.0389
	ξ_3	–0.0435	0.0622	–0.0708	0.0432
	α	–0.0099	0.0034	–0.0026	0.0033

response, where the true model is actually characterized by an ordinary single-index linear model, the proposed method would still provide consistent estimates, though some efficiency loss may incur.

On the other hand, if the true underlying model is partially linear but we adopt a standard second-order GEE (in this case, Prentice's (1988) approach) model, then the resulting estimators could be biased. The biases for $\boldsymbol{\xi}$ estimates are apparent in Table 1 when covariates u_{ij}'s are generated from a continuous distribution. We also observe that SGEE2 provides lower MSE for the estimators of the linear coefficient β, but the difference is not profound. However, lower MSE from SGEE2 is more apparent for the estimators of $\boldsymbol{\xi}$ and α, which is not surprising.

In Table 2, the two methods appear to provide similar results when the true underlying model is partially linear. This is perhaps due to the fact that the non-linear covariates are all discrete binary variables. As a consequence, the estimated single-index $\hat{\varphi}(\mathbf{u}'\boldsymbol{\xi})$ is not a smooth function. Hence our proposed method works simply as an ordinary GEE method for this type of data set.

6. Applications

6.1 Analysis of ICHS Data

Alfred Sommer and colleagues conducted a study (which we will refer to as the Indonesian Children's Health Study or ICHS) in the Aceh province of Indonesia to determine the causes and effects of vitamin A deficiency in pre-school children (Sommer, 1982). We present an analysis of infectious disease data on 250 Indonesian children, a subset of the cohort studied by Somer, Katz, and Tarwotjo (1983). The preschool children were examined up to six consecutive quarters for the presence of respiratory infection. There were 1,200 observations in total. We consider complete data for the first four visits from 548 pre-school children with or without the respiratory infection.

We focus on the question of whether vitamin A deficient children are at increased risk of respiratory infection, which is one of the leading causes of morbidity and mortality in children from the developing world. Such a relationship is plausible because vitamin A is required for the integrity of epithelial cells, the first line of defence against infection in the respiratory tract. Here the goal is to draw inferences on the change in respiratory infection status in the presence of intraperson correlation based on our proposed method. The model parameters were estimated and compared using the second order GEE (GEE2) approach of Prentice (1988) and our proposed semiparametric approach (SGEE2).

We set the binary response variable $Y_{it} = 1$ if the ith child suffers from reparatory infection at the tth visit, and 0 otherwise, for $i = 1, \ldots, 137$ and $t = 1, \ldots, 4$. The covariates of interest include "Xerop", which represents presence/absence (1 or 0) of xerophthalmia, an ocular manifestation of chronic vitamin A deficiency; "Time" represents time t; "Age" represents the baseline age in months (centered at 36); height for age is represented by "Height", as a percent of the National Center for Health Statistics (NCHS) standard (centered at 90%), which indicates long-term nutritional status.

The marginal mean response $p_{it} = E(Y_{it})$ is defined as a function of the covariates in the form

$$\text{logit}(p_{it}) = \beta_0 + \beta_1(\text{Xerop})_i + \varphi(\xi_1 \text{Time}_{it} + \xi_2 \text{Age}_i + \xi_3 \text{Height}_i) \tag{49}$$

where φ is an unknown function. Exchangeable association structure is considered here. We take the standard normal density as the kernel. The data-driven bandwidth h is used as discussed in Section 3.2.

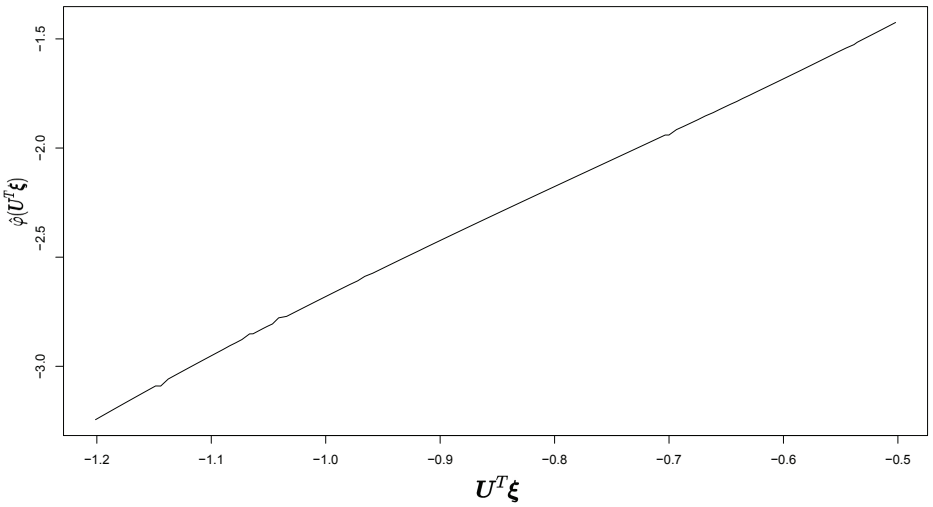

Figure 1. Estimated nonlinear curves for ICHS study.

Table 3. Estimates and Standard Errors (SEs) of Regression and Association Parameters for the ICHS Study.

Covariates	GEE2			SGEE2		
	Estimate	SE	z-value	Estimate	SE	z-value
Intercept	–0.734	0.761	–0.96	–0.249	0.762	–0.33
Xerophthalmia	1.448	0.634	2.29	1.451	0.628	2.31
Time	–1.452	0.641	–2.27	–0.525	0.264	–1.99
Age	–2.114	0.722	–2.93	–0.735	0.289	–2.54
Height for age	–1.133	0.788	–1.44	–0.429	0.308	–1.39
α	0.049	0.034	1.43	0.048	0.037	1.30

Figure 1 displays the estimates of the single index $\varphi(\mathbf{u}'\boldsymbol{\xi})$ against $\mathbf{u}'\boldsymbol{\xi}$. Here the estimated curve $\hat{\varphi}$ hardly shows any evidence of a nonlinear trend. This suggests that the data might be well fitted by an usual GEE model. Our finding agrees with that of Zeger and Karim (1991). In their study, the authors noted that there is no evidence that the height-for-age relationship deviates from the logistic-linear model for the most undernourished children. However, the inclusion of a nonlinear function φ allows model (49) to be more flexible to capture curvature, although the interpretation of the nonparametric covariate effects differs from that in an ordinary GEE model. In principle, a nonzero component of $\boldsymbol{\xi}$ suggests a significant predictor of the response, as commented in Carroll et al. (1997) .

Table 3 reports the estimates of the model parameters, their standard errors and corresponding z-values. We observe that the estimates of the linear covariate effects from the two methods are generally close to each other. The xerophthalmia coefficient is 1.448, indicating that the odds of suffering from respiratory infections is exp(1.448) = 4.25 times higher in vitamin A deficient children. The odds of suffering from respiratory infections appear to decrease with increased time, age and height for age. The presence of intraperson correlation appears to be very small and insignificant under both methods.

6.2 Analysis of Smoking Data

We also present an analysis of data on cigarette smoking trends from the Coronary Artery Development in Young Adults (CARDIA) study, an epidemiological study that recorded cardiovascular risk factors on five occasions over a 10-year period in black and white males and females (Hughes et al., 1987). This study was conducted in four urban centres (Birmingham, AL; Chicago, IL, Minneapolis, MN; and Oakland, CA) across the United States in which a total of 5,115 young adults aged 18-30 years were followed prospectively and examined up to five times from 1986 to 1996. Recruitment, restricted to blacks and whites, was carried out to achieve approximate balance in sample size with respect to age, race, gender, and education. Study participants were scheduled for visits at years 0, 2, 5, 7, and 10. We consider complete data for the first four visits from 3693 young adults with self reported smoking status (yes/no).

Here the goal is to draw inferences on the change in smoking prevalence of young adults in the presence of intraperson correlation based on our proposed method. The model parameters were estimated and compared using the second order GEE (GEE2) approach of Prentice (1988) and the proposed semiparametric approach (SGEE2).

Let the binary response variable $Y_{it} = 1$ if the ith individual is a smoker at the tth visit, and 0 if he/she is a nonsmoker. The marginal mean response $p_{it} = E(Y_{it})$ is defined as a function of the covariates in the form

$$\begin{aligned}\text{logit}(p_{it}) = \beta_0 + \beta_1(\text{Age}/10)_i + \beta_2\text{Time}_t + \beta_3\text{Eduh}_i + \beta_4\text{Educ}_i \\ +\varphi(\xi_1\text{Racebf}_i + \xi_2\text{Racewm}_i + \xi_3\text{Racewf}_i),\end{aligned} \tag{50}$$

for $i = 1, \ldots, 3693$ and $t = 1, \ldots, 4$, where Age_i = age of individual i in years at baseline time; Time_t = year since the baseline measurement = 0,2,5,7; the binary indicators $\text{Eduh}_i = 1$ if ith individual's education level is high school or less, and 0 otherwise; $\text{Educ}_i = 1$ if education level is up to some college, and 0 otherwise; $\text{Racebf}_i = 1$ if the person is a black female, and 0 otherwise; $\text{Racewm}_i = 1$ if the person is a white male, and 0 otherwise; and $\text{Racewf}_i = 1$ if the person is a white female, and 0 otherwise. Exchangeable association structure is modeled here. We take the standard normal density as the kernel. The data-driven bandwidth h is used as discussed in Section 3.2.

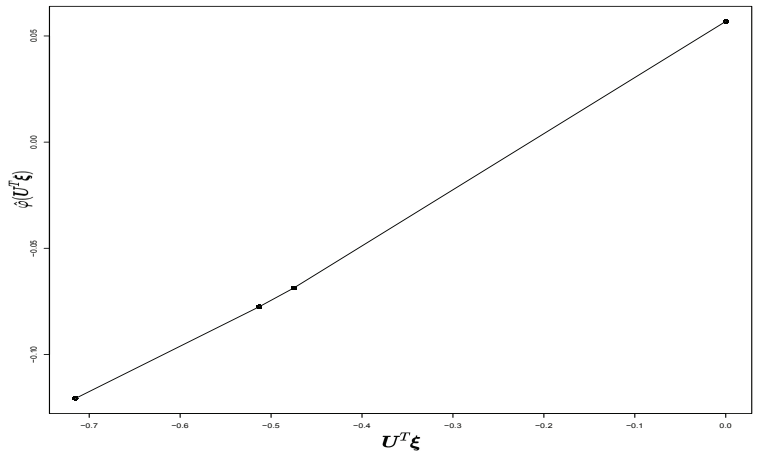

Figure 2. Estimated nonlinear curves for the smoking data from the CARDIA Study

Figure 2 displays the estimates of the single index $\varphi(\mathbf{u}'\boldsymbol{\xi})$ against $\mathbf{u}'\boldsymbol{\xi}$. Here the estimated curve $\hat{\varphi}(\mathbf{u}'\boldsymbol{\xi})$ does not show much evidence of a nonlinear trend, though a small curvature is visible. This suggests that the data might be well fitted by an ordinary GEE model.

Table 4 reports the estimates of the model parameters, their standard errors, and corresponding z-values. We observe that the estimates for all the covariates from both methods are generally close to each other. The primary reason for similar parameter estimates for the covariates is that the non-linear covariates that are included in this

Table 4. Estimates and Standard Errors (SEs) of Regression and Association Parameters for the CARDIA Study.

Covariates	GEE2			SGEE2		
	Estimate	SE	z-value	Estimate	SE	z-value
Intercept	–2.937	0.716	–4.10	–3.732	0.718	–5.20
Age/10	0.499	0.279	1.79	0.498	0.280	1.78
Time	–0.011	0.012	–0.92	–0.011	0.012	–0.92
Education (High School or less)	1.346	0.285	4.73	1.347	0.285	4.73
Education (Some college)	1.055	0.240	4.39	1.056	0.241	4.39
Race (Black Female)	–0.766	0.277	–2.77	–0.890	0.168	–5.30
Race (White Male)	0.046	0.281	0.16	–0.068	0.309	–0.22
Race (White Female)	–0.332	0.308	–1.08	–0.452	0.245	–1.84
α	0.741	0.030	24.95	0.730	0.058	12.63

model are all binary variables. As a result, our estimated single-index $\hat{\varphi}(\mathbf{u}'\boldsymbol{\xi})$ is a step function as we can see from Figure 2. Hence our proposed method performed as an ordinary GEE method for this data set.

The results from Table 4 suggest that the level of education has strong influence on the smoking status of the subjects. For example, the young adults are estimated to have exp(1.346) = 3.85 times higher odds to be a smoker if their level of education is up to high school or less than those who have a college degree or more. The presence of intraperson correlation appears to be high and significant under both methods.

7. Conclusions

The purpose of this research was to propose and explore a semiparametric approach to analyzing longitudinal binary data. Here the interest lies in the simultaneous estimation of the marginal mean parameters and association parameters. In some previous studies (e.g., Grace et al., 2009), an independent working correlation structure is considered when conducting estimation of the mean parameters. In our proposed method, we incorporate the true correlation structure into the estimating equations for the regression parameters.

The simulation results demonstrate that if the true underlying model is partially linear, then our proposed method generally provides unbiased and efficient estimators of the model parameters. Even if we consider a misspecified nonparametric function in the marginal mean response, which is actually characterized by an ordinary single-index linear model, the proposed method still leads to consistent estimates, although some efficiency loss may incur. Therefore, the proposed method has applications in a wide variety of settings. It can also be generalized for accommodating data with somewhat complex association structures.

Acknowledgements

This research is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*, 477–489. http://dx.doi.org/10.1080/01621459.1997.10474001

Diggle, P. J., Heagrety, P., Linang, K. Y., & Zeger, S. L. (2003). *Analysis of Longitudinal Data*. Oxford University Press Inc., New York.

Fan, J., & Li, R. (2004). New estimation and model selection procedures for semiparametric modelling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*, 710–723. http://dx.doi.org/10.1198/016214504000001060

Hastie, T. J., & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall: London.

Hughes, G. H., Cutter, G, Donahue, R., Freidman, G. D., Hully, S., Hunkeler, E., Jacobs, D. R., Liu, K., Orden, S.,Pirie, P., Tucker, B., & Wagenknecht, L. (1987). Recruitment in coronary artery risk development in young adults (CARDIA) study. *Controlled Clinical Trials*, *8*, 68S–73S. http://dx.doi.org/10.1016/0197-2456(87)90008-0

Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal o f the American Statistical Association*, *86*, 316-342. http://dx.doi.org/10.1080/01621459.1991.10475035

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22. http://dx.doi.org/10.1093/biomet/73.1.13

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized Linear, and Mixed Models*, second edition. Wiley, New Jersey.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, *9*(1), 141–142. http://dx.doi.org/10.1137/1109020

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics: Simulation and Computation*, *23*, 939–951. http://dx.doi.org/10.1080/03610919408813210

Preisser, J. S., Galecki, A. T., Lohman, K. K., & Wagenknecht, L. E. (2000). Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of the American Statistical Association*, *73*, 1021–1031. http://dx.doi.org/10.1080/01621459.2000.10474299

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, *44*, 1033–1048. http://dx.doi.org/10.2307/2531733

Watson, G. S. (1964). Smooth regression analysis, *The Indian Journal of Statistics*, Series A 26 (4), 359–372.

Yi, G. Y., He, W., & Liang, H. (2009). Analysis of correlated binary data under partially linear single-index logistic models. *Journal of Multivariate Analysis*, *100*, 278–290. http://dx.doi.org/10.1016/j.jmva.2008.04.012