Published in final edited form as:

Biometrics. 2013 June ; 69(2): 386–395. doi:10.1111/biom.12015.

Generalized Partially Linear Models for Incomplete Longitudinal Data In the Presence of Population-Level Information

Baojiang Chen¹ and Xiao-Hua Zhou²

¹Department of Biostatistics, College of Public Health, University of Nebraska Medical Center, Omaha, Nebraska, US 68198 baojiang.chen@unmc.edu

²Department of Biostatistics, University of Washington, Seattle, Washington, US 98195 azhou@u.washington.edu

Summary

In observational studies, interest often lies in estimation of the population-level relationship between the explanatory variables and dependent variables, and the estimation is often done using longitudinal data. Longitudinal data often feature sampling error and bias due to non-random drop-out. However, inclusion of population-level information can increase estimation efficiency. In this paper we consider a generalized partially linear model for incomplete longitudinal data in the presence of the population-level information. A pseudo-empirical likelihood-based method is introduced to incorporate population-level information, and non-random drop-out bias is corrected by using a weighted generalized estimating equations method. A three-step estimation procedure is proposed, which makes the computation easier. Several methods that are often used in practice are compared in simulation studies, which demonstrate that our proposed method can correct the non-random drop-out bias and increase the estimation efficiency, especially for small sample size or when the missing proportion is high. We apply this method to an Alzheimer's disease study.

Keywords

Auxiliary; drop-out; longitudinal data; partially linear model; population-level information; pseudo-empirical likelihood

1. Introduction

Longitudinal studies arise in many research areas such as medicine, public health and social science. The common feature of this kind of studies is the repeated measurements of the same subjects. During follow-up, subjects may drop out of the study due to, for example, their disease status. For instance, in the Alzheimer's disease (AD) study, at each clinic visit, a patient may drop out of the study due to developing dementia. Analyses based only on individuals with complete data can lead to invalid inferences if the mechanism leading to the missing data is dependent on the response. Under a missing completely at random (MCAR) mechanism (Little and Rubin, 2002), analyses based on generalized estimating equations (GEEs) (Liang and Zeger, 1986) yield consistent estimators of the regression parameters. However, when data are missing at random (MAR) or missing not at random (MNAR) (Little and Rubin, 2002), analyses based on GEEs generally give inconsistent estimators. Robins, Rotnitzky and Zhao (1995) developed a class of inverse probability weighted generalized estimating equations

Supplementary Materials

Supplementary materials, referenced in Section 3.4, are available with this paper at the Biometrics website on Wiley Online Library.

(IPWGEEs) which can yield consistent estimators when data are MAR. The weights are obtained from models for the missing data process, and these models must be correctly specified for the resulting estimators to be consistent.

To model the dependence of an outcome variable on covariates, it is often assumed that the mean of the response conforms a generalized linear model, where covariates pertaining to the response are presented in a linear form through a link function. In practice, the relationship between the response and covariates may be very complex and linear terms may not be adequate enough to feature that relationship. To allow for a more flexibility, there has been substantial recent interest in modeling covariate effects nonparametrically (e.g., Lin and Carroll, 2000). In many circumstances, a semiparametric generalized partially linear model (GPLM) is more desirable than modeling every covariate effect nonparametrically. This model assumes a linear term on some covariates $\mathbf{X}$ and a nonparametric term for a scalar covariate T . This model specification is appealing when the effects of $\mathbf{X}$ are of major interest and the effects of T are nuisance. This allows us to make inference of the effects $\mathbf{X}$ while making minimal assumptions on T . Marginal methods based on GEEs and their extensions have become intensively popular to address this kind of problem (e.g., Lin and Carroll, 2001a, b; Carroll et al., 1997; Wang, Carroll and Lin, 2005; Yi, He and Liang, 2009a, b).

Most approaches to studying GPLMs are focused on the complete data and only a few devote to incomplete data problem. Liang et al. (2004) studied a special case of the GPLMs, the partially linear model, with missing covariates. Liang (2008) extended it to the GPLMs with missing covariates. Wang and Sun (2007) studied the case of partially linear models with missing responses. To our knowledge, most studies on GPLMs have focused on cross-sectional studies and few devote to the longitudinal studies with missing data. One of the goals of this paper is to study the GPLMs in longitudinal data with dropout.

In practice, estimation of parameters is done using sample survey data since it is hard to collect all the required information for the whole population. To improve the estimation efficiency, sources of population-level data such as census data can be used, since these kind of data are comparatively free of sampling error and are often independent of the model parameters. Thus, incorporating this auxiliary information can lead to more accurate estimates and better inferences (Chaudhuri, Handcock and Rendall, 2008). For example, in the AD study introduced in Section 5, we are interested in investigating risk factors that influence the occurrence of the dementia in the target population, which is the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). A sample including 2324 subjects is used to study this problem. Besides this sample, auxiliary information, the proportions of dementia in the UDS are known for each year of follow up. For more details, please see Section 5. One of the goals of this study is to incorporate this auxiliary information to improve estimation efficiency. Chen and Qin (1993) adapted the empirical likelihood method (Owen, 2001) to incorporate available auxiliary information in finite population sampling. Chen and Sitter (1999) introduced a pseudo-empirical likelihood approach to the effective use of auxiliary information in complex surveys. Chaudhuri, Handcock and Rendall (2008) introduced a two-step estimation procedure for generalized linear models, incorporating population-level information via the empirical likelihood approach, which overcomes the numerical and implementation difficulties of direct constrained maximization. These articles focus on cross-sectional data. For longitudinal data, You, Chen and Zhou (2006) constructed a block empirical likelihood method using working independence; Xue and Zhu (2007) proposed another block empirical likelihood method through centering longitudinal data, in which they do not consider the within-subject correlation structure either; Wang, Qian and Carroll (2010) considered the element-wise and subject-wise empirical likelihood to incorporate the within-subject correlation structure of the repeated measurement, where they only consider the continuous response variable. All

literature on the empirical likelihood method for longitudinal data does not consider incorporating the auxiliary information and does not consider the missing data problem.

Recently, there has been considerable literature on analysis of missing data using the empirical likelihood method (e.g., Wang and Rao, 2002a, 2002b; Liang, Wang and Carroll, 2007; Qin and Zhang, 2007; Qin, Zhang and Leung, 2010). These methods mainly focus on generalized linear models in cross-sectional studies. For GPLMs, Liang and Qin (2008) considered the empirical likelihood-based inferences for partially linear models with missing covariates; Yang, Xue and Cheng (2009) studies the missing covariates problem for partially linear model using the empirical likelihood-based method. These articles only consider the partially linear models in cross-section studies with missing covariates and did not consider incorporating population-level information in the models either.

To our knowledge, little work has been devoted to GPLMs in longitudinal studies with missing data using the empirical likelihood-based method to incorporate population-level information. To fill up this gap, in this paper, we propose a pseudo-empirical likelihood-based method for the generalized partially linear models to incorporate the population-level information for longitudinal data with drop-out. This approach is appealing in that it not only extends the previous proposed methods for dealing with incomplete cross-sectional data to the incomplete longitudinal data, but it can improve estimation efficiency through incorporating the population-level information, especially for small sample size or when the missing proportion is high. When there is no population-level information, the proposed method reduces to the case of GPLMs in longitudinal studies with missing data, which is still a new research area. Also, the three-stage estimation method introduced in Section 3 makes the implementation much easier. Furthermore, the generalized partially linear model allows more flexible dependence of an outcome variable on covariates.

The remainder of this paper is organized as follows. In Section 2, we introduce notation and models. In Section 3, we give the forms of the estimating equations and provide details on estimation and inference. Simulation studies are described in Section 4. Data arising from an Alzheimer's disease study are analyzed in Section 5. Concluding remarks are made in Section 6.

2. Notation and models

Consider a sample involving n individuals in which each individual is to be examined at J assessment times. Let Y_{ij} denote the response for subject i at the j th assessment and $(\mathbf{X}_{ij}, T_{ij})$ denote the corresponding covariate vector, where $\mathbf{X}_{ij}$ is a $p \times 1$ vector and T_{ij} is a scalar. For convenience we let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T, \dots, \mathbf{X}_{iJ}^T)^T$, and $\mathbf{T}_i = (T_{i1}, \dots, T_{iJ})^T$, $i = 1, \dots, n$.

The conditional mean of Y_{ij} is denoted as $\mu_{ij} = E(Y_{ij}|\mathbf{X}_i, \mathbf{T}_i)$, and we let $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iJ})^T$ denote the full vector of means. We suppose the mean of Y_{ij} depends on the covariate vector for subject i at time j through a model of the form

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \theta(T_{ij}), \quad (1)$$

for $j = 1, \dots, J$, $i = 1, \dots, n$, where $g(\cdot)$ is a monotone differentiable link function, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients of interest, $\theta(\cdot)$ is an unknown smooth function. As commonly assumed in generalized linear model, the variance $\text{var}(Y_{ij}|\mathbf{X}_i, \mathbf{T}_i)$ only depends on the marginal mean μ_{ij} . Here, it is noted that we make the implicit assumption that $E(Y_{ij}|\mathbf{X}_i,$

$\mathbf{T}_i) = E(Y_{ij}|\mathbf{X}_{ij}, T_{ij})$ and $\text{var}(Y_{ij}|\mathbf{X}_i, \mathbf{T}_i) = \text{var}(Y_{ij}|\mathbf{X}_{ij}, T_{ij})$ as is commonly done (e.g., Lin and Carroll, 2000).

One challenge for this kind of study is the incomplete data problem. To indicate the availability of data we let $R_{ij} = 1$ if subject i is observed at time point j , and 0 otherwise. Here we consider a monotone missing pattern where $R_{ij} = 0$ indicates that $R_{ik} = 0$ for $k \geq j$. Furthermore, we consider a missing at random (MAR) mechanism; that is,

$$P(R_{ij}=1|R_{i,j-1}=1, Y_i, \mathbf{X}_i, \mathbf{T}_i) = P(R_{ij}=1|R_{i,j-1}=1, \bar{Y}_{i,j-1}, \bar{\mathbf{X}}_{i,j-1}, \bar{\mathbf{T}}_{i,j-1}), \quad (2)$$

where $\bar{Y}_{ij} = \{Y_{i1}, \dots, Y_{ij}\}$, $\bar{\mathbf{X}}_{ij} = \{\mathbf{X}_{i1}, \dots, \mathbf{X}_{ij}\}$, and $\bar{\mathbf{T}}_{ij} = \{T_{i1}, \dots, T_{ij}\}$. Let $\lambda_{ij} = P(R_{ij} = 1 | R_{i,j-1} = 1, \bar{Y}_{i,j-1}, \bar{\mathbf{X}}_{i,j-1}, \bar{\mathbf{T}}_{i,j-1})$, we then can build a generalized linear model for λ_{ij} , say, logistic regression, as follows:

$$\text{logit } \lambda_{ij} = \mathbf{u}_{ij}^T \boldsymbol{\alpha}, \quad (3)$$

where $\mathbf{u}_{ij}$ is a function of $\{\bar{Y}_{i,j-1}, \bar{\mathbf{X}}_{i,j-1}, \bar{\mathbf{T}}_{i,j-1}\}$, and $\boldsymbol{\alpha}$ is an unknown vector of parameters. Let $\pi_{ij} = \prod_{k=1}^j \lambda_{ik}$ be the marginal probability that subject i at time j is observed.

3. Pseudo-empirical likelihood for longitudinal data in the presence of population-level information

3.1 Pseudo-empirical likelihood

Let n_j denote the number of subjects who are observed up to (and including) j but missing after j , $j = 1, \dots, J$. For simplicity, we assume the first n_j subjects are observed at time j . Define p_{ij} the probability mass to subject i at time j for $j = 1, \dots, J$, and $i = 1, \dots, n$. We define a pseudo-empirical likelihood for the observed longitudinal data as

$$L = \prod_{j=1}^J \prod_{i=1}^{n_j} p_{ij}^{1/\pi_{ij}} = \prod_{j=1}^J \prod_{i=1}^n p_{ij}^{r_{ij}/\pi_{ij}}, \quad (4)$$

by assuming “ $0^0 = 1$ ”, where r_{ij} is a realization of R_{ij} . The motivation for (4) is from the inverse probability weighting method (Robins, Rotnitzky and Zhao, 1995). The idea behind this is that the weight has eliminated the bias, by “reconstructing” the log-empirical likelihood by upweighting the data from subjects who have a small chance of being observed. To be specific, by taking expectation with respect to the missing data probability to the log-pseudo-empirical

likelihood (4), we can obtain the log-empirical likelihood $\sum_{j=1}^J \sum_{i=1}^n \log(p_{ij})$, by noting that $E(r_{ij}/\pi_{ij}) = 1$ if the missing data model is correctly specified.

From the properties of a distribution function it follows that $p_{ij} \geq 0$ and $\sum_{j=1}^J \sum_{i=1}^n p_{ij} = 1$. However, with missing data, p_{ij} ($i > n_j$) are not estimable, so we must add constraints to the observed mass p'_{ij} s. Again, motivated by the inverse probability weighting method, we upweight the data from subjects who have a small chance of being observed. Therefore, we would rather use the following constraints

$$p_{ij} \geq 0, \sum_{j=1}^J \sum_{i=1}^n \frac{r_{ij}}{\pi_{ij}} p_{ij} = 1. \quad (5)$$

3.2 Population-level constraints

Suppose that there is a known vector function $\mathbf{h}_{ij}$ at each time point j , for which $E_j\{\mathbf{h}_{ij}(Y_{ij}, \mathbf{X}_{ij}, \mathbf{A}_j)\} = \mathbf{0}$ with known parameter $\mathbf{A}_j, j = 1, \dots, J$. Knowledge of $\mathbf{A}_j$ represents the population-level information at time point j . For example, in some survey study, we often know the population mean at each time point from census data, then $h_{ij} = Y_{ij} - A_j$, where A_j is the population mean at time point j . With the population-level information, we can maximize the pseudo-empirical likelihood (4) subject to the constraints (5) and

$$\sum_{i=1}^n \frac{r_{ij}}{\pi_{ij}} p_{ij} \mathbf{h}_{ij}(Y_{ij}, \mathbf{X}_{ij}, \mathbf{A}_j) = \mathbf{0}, j = 1, \dots, J. \quad (6)$$

The motivation of using the weighted constraint (6) is still based on the inverse probability weighting method by upweighting the data from subjects who have a small chance of being observed.

3.3 Generalized partially linear models with missing data

In this section we describe a marginal method for estimating parameters β and the nonparametric function $\theta(\cdot)$. Let $\mathbf{W}_i = \mathbf{B}_i^{1/2} \mathbf{C}_i \mathbf{B}_i^{1/2}$ be a working covariance matrix, where $\mathbf{B}_i = \text{diag}(v_{ij}(T_{ij}), j = 1, \dots, J)$, $v_{ij}(T_{ij}) = \text{var}(Y_{ij} | \mathbf{X}_{ij}, T_{ij})$, and $\mathbf{C}_i$ is an invertible working correlation matrix. Let $\mathbf{U}_i(\beta, \theta(\cdot)) = \partial \mu_i^T / \partial \beta \Delta_i \mathbf{W}_i^{-1} (Y_i - \mu_i) = 0$, where $\Delta_i = \text{diag}(r_{ij} / \pi_{ij} p_{ij}, j = 1, \dots, J)$. It can be seen that $\mathbf{U}_i(\beta, \theta(\cdot))$ has zero expectation if the missing data model and the marginal mean of the response are correctly specified.

Since the estimating function $\mathbf{U}_i(\beta, \theta(\cdot))$ involves an unknown smooth function $\theta(\cdot)$, we need to use nonparametric approaches to estimate this function locally. Assuming $\theta(u)$ has the second derivative, we may approximate $\theta(u)$ by a locally linear function within the neighborhood of u_0 via the Taylor series expansion $\theta(u) \approx \theta(u_0) + \theta'(u_0)(u - u_0)$ for a given point u_0 . Denote $a_0(u_0) = \theta(u_0)$, $a_1(u_0) = \theta'(u_0)$, and $\mathbf{a}(u_0) = (a_0(u_0), a_1(u_0))^T$. Let $\mathbf{T}_{ij}(u) = (1, T_{ij} - u)^T$, $\mathbf{T}_i(u) = (\mathbf{T}_{i1}^T(u), \dots, \mathbf{T}_{ij}^T(u))^T$. Let $K(u)$ be a kernel function with a compact support and h be a bandwidth. Denote $K_h(t) = K(t/h)/h$ and $K_{ih}(u) = \text{diag}(K_h(T_{ij} - u), j = 1, \dots, J)$. Let $\mathbf{D}_i(\mathbf{T}_i) = \text{diag}(\mu_{ij}^{(1)}, j = 1, \dots, J)$, where $\mu_{ij}^{(1)}$ denotes the first derivative of $\mu_{ij}(\cdot)$ evaluated at $\mathbf{X}_{ij}^T \beta + \theta(T_{ij})$.

Below we describe a three-stage estimation procedure for estimation of the parameter β and the nonparametric function θ in the presence of incomplete data and population-level information, as follows:

Stage 1. Estimate the missing data parameter α via a maximum likelihood method, then calculate the estimated weights $\hat{\pi}_{ij}$.

Stage 2. Maximize likelihood (4) on the simplex, under the constraints (5) and (6), with respect to p_{ij} . Denote the estimated value $\hat{p}_{ij}$.

Stage 3. Solve the parameter β and the nonparametric function $\theta(\cdot)$ via the following steps:

- a. Choose initial value $\beta^{(0)}$, set $\widehat{\beta} = \beta^{(0)}$.
- b. For a given point u in a selected grid, find $\widehat{\theta}(u; \widehat{\beta}) = \widehat{a}_0(u)$ by solving the following estimating equations

$$\sum_{i=1}^n \mathbf{T}_i^T(u) \widetilde{\mathbf{D}}_i(u) \widehat{\Delta}_i K_{ih}(u) \widetilde{\mathbf{W}}_i(u)^{-1} (\mathbf{Y}_i - \widetilde{\boldsymbol{\mu}}_i) = 0, \quad (7)$$

with respect to $\mathbf{a}(u)$, where $\widetilde{\boldsymbol{\mu}}_i = (\widetilde{\mu}_{i1}, \dots, \widetilde{\mu}_{ij})^T$ with $\widetilde{\mu}_{ij} = g^{-1}(\mathbf{X}_{ij}^T \widehat{\beta} + \mathbf{T}_{ij}^T(u) \mathbf{a}(u))$, $\widetilde{\mathbf{D}}_i(u)$ is $\mathbf{D}_i(u)$ with μ_{ij} replaced by $\widetilde{\mu}_{ij}$, and $\widetilde{\mathbf{W}}_i(u)$ is the independence working covariance matrix $\text{diag}(\widetilde{v}_{ij}(u), j=1, \dots, J)$, $\widetilde{v}_{ij}(u)$ is $v_{ij}(u)$ with μ_{ij} replaced by $\widetilde{\mu}_{ij}$, and $\widehat{\Delta}_i = \text{diag}(\widehat{N} r_{ij} \widehat{\pi}_{ij} \widehat{p}_{ij}, j=1, \dots, J)$ with $\widehat{N} = \sum_{i=1}^n \sum_{j=1}^J r_{ij} \widehat{\pi}_{ij}$.

- c. given the estimate $\widehat{\theta}(T_{ij}; \widehat{\beta}) = \widehat{a}_0(T_{ij})$ for points T_{ij} , update $\widehat{\beta}$ by solving equations for β :

$$\sum_{i=1}^n \frac{\partial \widehat{\boldsymbol{\mu}}_i(\beta)}{\partial \beta^T} \widehat{\Delta}_i \mathbf{W}_i^{-1} (\mathbf{Y}_i - \widehat{\boldsymbol{\mu}}_i(\beta)) = 0,$$

where $\widehat{\boldsymbol{\mu}}_i(\beta) = (\widehat{\mu}_{i1}(\beta), \dots, \widehat{\mu}_{ij}(\beta))^T$ with $\widehat{\mu}_{ij}(\beta) = g^{-1}(\mathbf{X}_{ij}^T \beta + \widehat{\theta}(T_{ij}; \widehat{\beta}))$, and $\mathbf{W}_i = \text{diag}(\widehat{v}_{ij}(T_{ij}), j=1, \dots, J)$, $\widehat{v}_{ij}(T_{ij})$ is $v_{ij}(T_{ij})$ with μ_{ij} replaced by $\widehat{\mu}_{ij} = g^{-1}(\mathbf{X}_{ij}^T \beta + \widehat{\theta}(T_{ij}; \widehat{\beta}))$.

- d. Iterate steps (b) and (c) until convergence.

It is easy to verify that there is a unique maximizing $\widehat{p} = (\widehat{p}_{ij}, j=1, \dots, J, i=1, \dots, n_j)$ under the constraints (5) and (6). Empirical evidence suggests that the three-stage procedures work well.

3.4 Estimation and inference

The first stage involves estimating the missing data model parameter α . This can be done by maximizing the log-likelihood for the parameter α :

$$\sum_{i=1}^n \ell_i(\alpha) = \sum_{i=1}^n \left\{ \sum_{j=1}^{t_i-1} \log(\lambda_{ij}) + \log(1 - \lambda_{it_i}) \right\},$$

where $t_i = \min\{j : R_{ij} = 0\}$. Equivalently, we solve the estimating equation $\sum_{i=1}^n \mathbf{S}_i(\alpha) = 0$, where $\mathbf{S}_i(\alpha) = \partial \ell_i(\alpha) / \partial \alpha^T$. This procedure can be done by using the general software. Denote the estimator $\widehat{\alpha}$. Thus we can obtain the estimated marginal probability $\widehat{\pi}_{ij}$.

Note that the second stage requires constrained maximization. For computational purposes, we maximize the log-empirical likelihood

$$\ell_r = \sum_{j=1}^J \sum_{i=1}^n \frac{r_{ij}}{\pi_{ij}} \log(p_{ij}), \quad (8)$$

subject to the constraints (5) and (6). Based on the dual maximization problem (Owen, 2001), we have $p_{ij} = r_{ij} / \{N(1 + \eta_j^T \mathbf{h}_{ij})\}$, $j = 1, \dots, J$, $i = 1, \dots, n_j$, where $N = \sum_{i=1}^n \sum_{j=1}^J r_{ij} / \pi_{ij}$, η_j is a vector of the Lagrange multipliers determined by the equations

$$\sum_{i=1}^n \mathbf{H}_{ij}(\eta, \alpha) = 0, \quad j = 1, \dots, J, \quad (9)$$

where $\mathbf{H}_{ij}(\eta, \alpha) = r_{ij} \mathbf{h}_{ij} / [N \pi_{ij} (1 + \eta_j^T \mathbf{h}_{ij})]$. Let $\mathbf{H}_i(\eta, \alpha) = (\mathbf{H}_{i1}^T(\eta, \alpha), \dots, \mathbf{H}_{iJ}^T(\eta, \alpha))^T$.

Let $\eta = (\eta_1^T, \dots, \eta_J^T)^T$, and $\widehat{\eta} = (\widehat{\eta}_1^T, \dots, \widehat{\eta}_J^T)^T$ be the estimator which solves estimating equations (9). Then the estimated probability $\widehat{p}_{ij} = r_{ij} / \{\widehat{N}(1 + \widehat{\eta}_j^T \mathbf{h}_{ij})\}$, $j = 1, \dots, J$, $i = 1, \dots, n_j$.

Replacing $\widehat{\Delta}_i = \text{diag}(\widehat{p}_{ij} r_{ij} / \pi_{ij}, j = 1, \dots, J)$ in the third stage and following (a)-(d), we can obtain the estimator of β , say $\widehat{\beta}$, and the nonparametric function $\widehat{\theta}(t)$.

We now develop the asymptotic properties of $\widehat{\beta}$ and $\widehat{\theta}(t)$.

THEOREM 1: Suppose that $(Y_1, \mathbf{X}_1, T_1), \dots, (Y_n, \mathbf{X}_n, T_n)$ are independent and identically distributed random vectors from the population space governed by model (1). Assuming the regularity conditions stated in the Appendix in the Supplementary Materials hold, if the missing data parameter α and the Lagrange multiplier η are known, and as $nh^2/\log(1/h) \rightarrow \infty$, we have

(a) The estimate of $\theta(t)$, $\widehat{\theta}(t; \widehat{\beta})$, has the following property

a. The estimate of $\theta(t)$, $\widehat{\theta}(t; \widehat{\beta})$, has the following property

$$(nh)^{1/2} \{\widehat{\theta}(t; \widehat{\beta}) - \theta(t) - \theta^{(2)}(t) h^2 / 2\} \rightarrow N(0, \gamma W_2(t)^{-1}),$$

where $\gamma = \int K(u)^2 du$, $W_2(t) = \sum_j E[\{\mu_j^{(1)}\}^2 v_j^{-1} | T_j = t] f_j(t)$, $v_j = \text{var}(Y_j | \mathbf{X}_j, T_j)$, $f_j(t)$ is the marginal density of T_j , and $\theta^{(2)}(t)$ is the second derivative of $\theta(t)$.

b. $n^{1/2}(\widehat{\beta} - \beta) \rightarrow N(0, \Gamma^{-1} \Sigma [\Gamma^{-1}]^T)$, as $n \rightarrow \infty$, where $\Sigma = E[\mathbf{Q}_i \mathbf{Q}_i^T]$, and Γ and $\mathbf{Q}_i$ are given in the Appendix in the Supplementary Materials.

In practice, the missing data parameter, hence the missing data probabilities are unknown and we must estimate them consistently. As recommended at the beginning of this subsection, we may employ the maximum likelihood estimate. In this case, we have the following theorem.

THEOREM 2: Suppose that $(Y_1, \mathbf{X}_1, T_1), \dots, (Y_n, \mathbf{X}_n, T_n)$ are independent and identically distributed random vectors from the population space governed by model (1). Assuming the regularity conditions stated in the Appendix in the Supplementary Materials hold, the missing data model is correctly specified, and as $nh^4 \rightarrow 0$ and $nh^2/\log(1/h) \rightarrow \infty$, we have

$$n^{1/2}(\widehat{\beta} - \beta) \rightarrow N\left(0, \Gamma^{-1} \Sigma^* [\Gamma^{-1}]^T\right),$$

as $n \rightarrow \infty$, where $\Sigma^* = E[\mathbf{Q}_i^* \mathbf{Q}_i^{*T}]$, and Γ and $\mathbf{Q}_i^*$ are given in the Appendix in the Supplementary Materials.

Along with the proof in the Appendix, we can derive an approximation of $\widehat{\theta}(t; \widehat{\beta}, \widehat{\alpha}, \widehat{\eta}) - \theta(t)$, where $\widehat{\theta}(t; \widehat{\beta}, \widehat{\alpha}, \widehat{\eta})$ is the estimate of $\theta(t)$ by profiling the estimate $\widehat{\beta}$, $\widehat{\alpha}$ and $\widehat{\eta}$. Standard calculations can show that $(nh)^{1/2} \{\widehat{\theta}(t; \widehat{\beta}, \widehat{\alpha}, \widehat{\eta}) - \theta(t) - \theta^{(2)}(t)h^2/2\}$ weakly converges to a normal distribution with mean zero. This result in Theorem 2 may be used for statistical inference. However, we suggest a bootstrap alternative if needed since the asymptotic variance has a very complex form.

Theorems 1 and 2 indicate that in order to estimate β at the root- n rate, one must undersmooth the nonparametric part $\theta(t)$, which is standard in the kernel literature.

As is widely known, bandwidth h affects both bias and variance estimate, and there is a trade-off between suitable bias and variance estimate. The choice of a relatively small bandwidth can correct biases, whereas the choice of a large value of bandwidth can yield a smaller variance. In principle, bandwidth selection is data driven, and tradition methods such as cross-validation approach may be applied to select a proper bandwidth h based on available data (Yi, He and Liang, 2009a). However, Fan, Heckman and Wand (1995) pointed out that this approach could perform poorly in some settings with a large magnitude of sample variation produced, hence it is not regarded as a sensible bandwidth selection rule in practice. As an alternative, the “plugging in” method may be a promising candidate. Rubbert, Scheather and Wand (1995) explored this method for local least squares regression. Fan, Heckman and Wand (1995) discussed this approach to handle local polynomial regression under the frame work of generalized linear models.

In the same spirit, an optimal bandwidth is often chosen based on the asymptotic weighted mean integrated square error (AMISE) of $\widehat{\theta}(\cdot; h)$,

$$AMISE\{\widehat{\theta}(\cdot; h)\} = E\left[\int \{\widehat{\theta}(u; h) - \theta(u)\}^2 f(u) du\right].$$

Carroll et al. (1997) discussed bandwidth selection using the criterion of minimizing AMISE for generalized linear single index models. For a given function $\omega(u)$ with compact support, at the final step of estimation of $\theta(u)$, minimizing the AMISE with weight $f(u)\omega(u)$ leads to the optimal bandwidth

$$h_{opt} = n^{-1/5} \left\{ \frac{\int K^2(t) dt}{\left(\int t^2 K(t) dt\right)^2} \right\} \left\{ \frac{d_*(u) \omega(u)}{\int \theta^{(2)}(u)^2 f(u) \omega(u) du} \right\}^{1/5},$$

with some form $d_*(u)$. Plug-in bandwidth strategies may be applied to replace the unknown integrals by their approximations. As noted in Carroll et al. (1997), a sensible choice of

bandwidth h is generally difficult. Instead, Carroll et al. (1997) suggested an ad hoc bandwidth, given by $\widehat{h}_{opt} \times n^{-2/15} = O(n^{-1/5} \times n^{-2/15}) = O(n^{-1/3})$.

4. Simulations

In this section, we assess the empirical performance of the proposed methods through the simulation studies, and contrast their performance with that of other methods that may be commonly used in practice.

We consider a binary longitudinal study with $J = 3$, where we suppose there is a superpopulation with 10000 subjects, which is generated from model

$$\text{logit } \mu_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + \theta(T_{ij}),$$

where $\mu_{ij} = E(Y_{ij}|X_{1ij}, X_{2ij}, T_{ij})$, $\theta(t) = \sin(2t)$, X_{1ij} and X_{2ij} are time varying covariates. We generate X_{1ij} , X_{2ij} and T_{ij} according to models $X_{1ij} = b_i + e_{1ij}$, $X_{2ij} = b_i + e_{2ij}$, $T_{ij} = b_i + e_{3ij}$, where $b_i \sim \text{uniform}(-1, 1)$, and e_{1ij} , e_{2ij} and e_{3ij} are independent and follow $\text{uniform}(-1, 1)$. This setup allows X_{1ij} , X_{2ij} and T_{ij} to be correlated with each other. For the response process, we assume the true correlation matrix over time is exchangeable with correlation coefficient ρ . The true values of coefficients are $\beta_0 = \log(1.5)$, and $\beta_1 = \log(0.5)$. The study sample of size n is randomly sampled from the superpopulation.

The missing data model is

$$\text{logit } \lambda_{ij} = \alpha_0 + \alpha_1 Y_{i,j-1}, \quad (10)$$

where $\lambda_{ij} = P(R_{ij} = 1 | R_{i,j-1} = 1, Y_{i,j-1})$, and the true value $\alpha_0 = 3$; we take α_1 as -5 and -2 , leading to the missing proportion 35% and 15% respectively. We assume all subjects are observed at time 1.

Suppose we know the population mean at each time point (obtained from the superpopulation), denoted as A_j , $j = 1, \dots, J$; the constraints are

$$\sum_{i=1}^n \frac{r_{ij}}{\pi_{ij}} p_{ij} (Y_{ij} - A_j) = 0, \quad j = 1, 2, 3.$$

Here we assess the performance of the proposed method along with other methods that might be used in practice. Specifically, we compare the following methods: (a) Method 1 is the proposed method; (b) Method 2 is the proposed method without considering the population-level information, i.e. we use the weight matrix $\Delta_i^w = \text{diag}(r_{ij}/\pi_{ij}, j = 1, 2, 3)$ instead of using the weight matrix $\Delta_i = \text{diag}(r_{ij}p_{ij}/\pi_{ij}, j = 1, 2, 3)$; (c) Method 3 is the generalized linear model:

$$\text{logit } \mu_{ij} = \beta_1^* X_{1ij} + \beta_2^* X_{2ij} + \beta_3^* T_{ij}, \quad (11)$$

where we use the weight matrix Δ_i and the true working correlation matrix; (d) Method 4 is the method of Robins, Rotnitzky and Zhao (1995) by fitting the model (11) using the weight matrix Δ_i^w and the true working correlation matrix. The kernel function used in Methods 1 and 2 is the Gaussian kernel function, i.e. $K(x) = \phi(x)$, where $\phi(x)$ is the standard normal density function. The bandwidth is chosen as $h = 2 \times n^{-1/3}$.

Table 1 lists the results for parameters β_1 and β_2 with correctly specified missing data model (10), where bias is the percent relative bias, SD is the standard deviation for the 2000 simulations, and CP is the percent coverage probability for the nominal level 95% in which the standard errors are estimated using the bootstrap method based on 1000 bootstrap samples. It can be seen that Methods 1 and 2 give consistent estimates and good coverage probabilities under all situations, but Methods 3 and 4 give biased estimates. Method 1 is more efficient than Method 2 for all the cases, especially when the missing proportion is high, indicating that incorporating auxiliary information will yield more efficient estimates for the parametric parameters. The efficiency gain depends on the sample size, the missing proportion, and the response correlation ρ . As the missing proportion increases (α_1 decreases), the efficiency gain increases; as ρ increases, the efficiency gain increases.

Figures 1 and 2 compare Methods 1 and 2 through the biases and efficiency gain for the nonparametric function $\theta(t)$, where we assume $n = 400$, $\rho = 0.6$, and $\alpha_1 = -5$. Bootstrap methods are used for estimating the variance of $\hat{\theta}$. Both estimates of $\theta(t)$ are close to the true $\theta(t)$; the proposed method (Method 1) yields more efficient estimate of $\theta(t)$ than that based on Method 2, indicating that incorporating auxiliary information will yield more efficient estimates of the nonparametric function.

We comment that the proposed method runs fast. For example, with a general personal computer using software R 2.9.2, the CPU time is 0.11 and 0.06 second for Methods 1 and 2 when $n = 400$, $\rho = 0.6$ and $\alpha_1 = -2$; the average CPU time is 0.16 and 0.08 second for Methods 1 and 2 for each run when $n = 400$, $\rho = 0.6$ and $\alpha_1 = -5$. With a smaller sample size $n = 200$, the average CPU time is 0.09 and 0.05 second for Methods 1 and 2 for each run when $\rho = 0.6$ and $\alpha_1 = -2$; and the average CPU time is 0.14 and 0.07 second for Methods 1 and 2 for each run when $\rho = 0.6$ and $\alpha_1 = -5$.

Table 1 also compares the estimates of all the methods where the true model is a linear model:

$$\text{logit } \mu_{ij} = \beta_1 X_{1ij} + \beta_2 X_{2ij} + T_{ij}.$$

It can be seen that all the methods yield consistent estimates and good coverage probabilities; Method 1 gives more efficient estimates than Method 2, and Method 3 give more efficient estimate than Method 4; when the missing proportion is high (α_1 is small), Method 1 yields more efficient estimates than Method 3; when the missing proportion is low (α_1 is big), Method 1 yields a little bit less efficient estimates than Method 3, but the efficiency loss is small. Plotting $\theta(t)$ (not shown here) also demonstrates that Method 1 yields consistent estimate for the nonparametric function. All indicate that the proposed method is safe to use, even when the model is linear.

Finally, to assess the performance of the proposed method to the misspecification of the missing data model, we considering fitting the missing data model

$$\text{logit } \lambda_{ij} = \alpha_0 + \alpha_1 T_{i,j-1},$$

while the true model is (10). Results are shown in Table 2. As is expected, estimation of β is sensitive to misspecification of the model for the missing data process when $\rho \neq 0$. When $\rho = 0$, it seems that Methods 1 and 2 are robust to the misspecification of the missing data model in this simulation study. Therefore, careful assessment of the missing data model is warranted. An appealing approach is to assess a particular model by comparing its fit to expanded models

including additional terms. This provides focused tests of the adequacy of a particular model which are easily implemented.

5. Application to an Alzheimer's disease study

We apply the proposed method to the National Alzheimer's Coordinating Center (NACC) Uniform Data Set (UDS). One of the goals of the study is to investigate the risk factors that influence the occurrence of dementia in this population, which consists of all patients in 31 NIA funded Alzheimer's disease centers. The response of interest is the clinical diagnosis for dementia (Yes/No), because definitive diagnosis of dementia is possible only through neuropathological examination, which is not available for the majority of subjects. The covariates that may influence the occurrence of dementia include sex, family history of dementia (FHDEM, yes/no), diabetes (yes/no), behavioral assessment (depression or dysphoria, yes/no), hypertension (yes/no), education (years), and age, in which diabetes, behavioral assessment, and hypertension are time-varying, and the others are not.

A sample of 2324 subjects enrolled in 2005 in the UDS was used to study the relationship between covariates and responses using the population-based information in the present rates of dementia. Follow-up visits for subjects are scheduled at approximately one-year intervals, with up to four clinical visits until 2008. The auxiliary data is the UDS population, and we know the proportions of dementia in the UDS population from 2005 to 2008, which are 69.6%, 63.1%, 63.1%, 66.4%, respectively. We proposed to use this sort of auxiliary information to improve estimation efficiency. During the study, patients may drop out the study. We think the MAR mechanism may be reasonable because, for example, a patient's visit to a clinic may depend on his/her previous observed diagnosis of disease status: if s/he was diagnosed dementia last year, s/he may be likely to attend the clinic visit this year in order to control the disease efficiently. The missing proportion is about 35.5%.

Consider the following regression model for the response process

$$\text{logit } \mu_{ij} = \mathbf{u}_{ij}^T \beta + \theta(\text{age}_{ij}), \quad (12)$$

where $\mathbf{u}_{ij}$ is the covariate vector at time point j , which includes the function of sex, DEPD, FHDEM, diabetes, depression, hypertension, education, and the indicator variables for centers.

For the drop-out process, we build the following regression models:

$$\text{logit } \lambda_{ij} = \mathbf{v}_{ij}^T \alpha,$$

where $\mathbf{v}_{ij}$ include functions of sex, DEPD, FHDEM, diabetes, hypertension, education, age, and previous response.

In line with the simulation study, here we also use four methods to analyze the data. The first method is the proposed method based on the weight matrix Δ_i ; the second method is the proposed method based on the weight matrix Δ_i^w ; the third method is the generalized estimating equation method based on Robins, Rotnitzky and Zhao (1995) where age is considered as a linear predictor, and the weight matrix is Δ_i ; the fourth method is the generalized estimating equation method where age is considered as a linear predictor, and the weight matrix is Δ_i^w ; the results are reported in Table 3. Here we only list the effects for the risk factors and omit the center effects though some of them are highly significant. It can be seen that Methods 1 and 2 give close results, and they give different results from Methods 3 and 4, since the nonparametric function $\theta(t)$ is nonlinear (see Figure 3), so Methods 3 and 4 will give biased results. Method

1 yields smaller standard errors than those for Method 2, indicating that incorporating the population-level information yields more efficient estimates. All methods reveal that SEX, depression, diabetes and hypotension have significant effect on the occurrence of dementia; family history of dementia and education have no significant effects on the occurrence of dementia.

Figure 3 plots the nonparametric function $\hat{\theta}(\text{age})$. It indicates that the age effect departs a little from linearity, especially at the tail, so the semiparametric model (12) may be reasonable than the logistic linear model. However, we should note that the confidence intervals at the tail of Figure 3 are wide. This is because of the small sample size, which inflates the standard errors. Hence, for simplicity, a linear function can be used for the pattern of $\theta(\text{age})$. Here, we suggest using the semiparametric model (12) if the age effect is not of interest, as it is much safer to use no matter the true function form of $\theta(\text{age})$ is linear or not (see discussion in Section 4.)

For the missing data model, we carry out standard diagnostic tests for the fit of regression models by comparing a model with an expanded model to do a model selection. Here, we only list the results for the final model in Table 3. Significance of sex, DEPD, diabetes, education, age, and the previous diagnosis of dementia, indicating that the missing at random mechanism may be feasible. So, we can draw the conclusions that the complete case analysis can draw a wrong conclusion about a relationship between a risk factor and the occurrence of dementia.

We remark that, to improve estimation efficiency, relevant auxiliary information should be used. In this example, we want to investigate the risk factors that influence the occurrence of dementia in this UDS population, so the auxiliary information (the proportions of dementia in the UDS population) is relevant. As is seen from the above analysis, incorporating this sort of information will increase estimation efficiency. In practice, not all sort of auxiliary information is useful to improve estimation efficiency. For example, simulation studies demonstrate that the marginal distribution of the outcome (such as mean and median of outcome) will help if we want to study the association between the outcome and covariates, however the marginal distribution of the covariates will provide little improvement in efficiency.

6. Discussion

In this paper, we propose an empirical likelihood-based method to incorporate the population-level information in a longitudinal study using the generalized partially linear model with non-random drop-out. The method has been demonstrated to be more efficient than the method that ignoring the population-level information, especially when the sample size is small or when the missing proportion is high. Also, incorporating the nonparametric function in the model makes it more flexible in practice. Simulation studies demonstrated that the proposed method works well in varieties of situations.

We focussed here primarily on estimation and inference regarding drop-out. Intermittently missing data are very common in practice. Future research is planned to extend this method to the intermittently missing data problem, with missing response and/or missing covariates. This is a challenging problem and will be under investigation in the future.

Note that, we only consider a parametric model in the missing data model. Similarly, we can consider the generalized partially linear model in the same spirit when some covariates are linear and some are non-linear. This is a natural extension of the proposed method and will not discuss it here. This modeling may correct the biases and/or improve the efficiency of the proposed method estimates. However, incorporating the nonparametric part will increase the computation burdensome. A trade-off is that we build a parametric missing data model, as has been done in this paper.

This paper only considers the missing at random mechanism. In practice, we often face data that are missing not at random. Another future research direction will be to extend this method to the missing not at random mechanism.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the Editor, the associated editor and two referees for their constructive comments and suggestions. Dr. Zhou's work was supported in part by U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D grants (RCS 05-196), and the National Science Foundation of China (NSFC 30728019). Both Drs. Zhou and Chen were support in part by National Institute on Aging grant U01AG016976.

References

- Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association*. 1997; 92:477–489.
- Chaudhui S, Handcock MS, Rendall MS. Generalized linear models incorporating population level information: an empirical-likelihood-based approach. *Journal of the Royal Statistical Society Series B*. 2008; 70:311–328.
- Chen J, Qin J. Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*. 1993; 80:107–116.
- Chen J, Sitter RR. A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistic sinica*. 1999; 9:385–406.
- Fan J, Heckman NE, Wand MP. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*. 1995; 90:141–150.
- Liang H. Generalized partially linear models with missing covariates. *Journal of Multivariate Analysis*. 2008; 99:880–895. [PubMed: 19412320]
- Liang H, Qin Y. Empirical likelihood-based inferences for partially linear models with missing covariates. *Australian & New Zealand Journal of Statistics*. 2008; 50:347–359. [PubMed: 20703365]
- Liang H, Wang S, Robins J, Carroll RJ. Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*. 2004; 99:357–367.
- Liang H, Wang S, Carroll RJ. Partially linear models with missing response variables and error-prone covariates. *Biometrika*. 2007; 94:185–198. [PubMed: 20209027]
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Lin X, Carroll RJ. Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*. 2000; 95:520–534.
- Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*. 2001a; 96:1045–1056.
- Lin X, Carroll RJ. Semiparametric regression for clustered data. *Biometrika*. 2001b; 88:1179–1185.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd ed.. John Wiley; New York: 2002.
- Owen, A. *Empirical likelihood*. Chapman and Hall-CRC; Boca Raton: 2001.
- Qin J, Zhang B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society Series B*. 2007; 69:101–122.
- Qin J, Zhang B, Leung DHY. Empirical likelihood in missing data problem. *Journal of the American Statistical Association*. 2010; 104:1492–1503.
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90:106–121.
- Rubbert D, Scheather SJ, Wand MP. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*. 1995; 90:1257–1270.

- Wang N, Carroll RJ, Lin X. Efficient semiparametric marginal estimation for longitudinal/clustering data. *Journal of the American Statistical Association*. 2005; 100:147–157.
- Wang S, Qian L, Carroll RJ. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*. 2010; 97:79–93. [PubMed: 20305730]
- Wang Q, Rao JNK. Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*. 2002a; 30:896–924.
- Wang Q, Rao JNK. Empirical likelihood-based inference in linear errors-in-covariates models with validation data. *Biometrika*. 2002b; 89:345–358.
- Wang Q, Sun Z. Estimation in partially linear models with missing responses at random. *Journal of Multivariate Analysis*. 2007; 98:1470–1493.
- Xue L, Zhu L. Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*. 2007; 94:921–937.
- Yang Y, Xue L, Cheng W. Empirical likelihood for a partially linear model with covariate data missing at random. *Journal of Statistical Planning and Inference*. 2009; 139:4143–4153.
- Yi GY, He W, Liang H. Analysis of correlated binar data under partially linear single index logistic models. *Journal of Multivariate Analysis*. 2009a; 100:278–290. [PubMed: 20126288]
- Yi GY, He W, Liang H. Semiparametric marginal and association regression methods for clustered binary data. *Annals of the Institute of Statistical Mathematics*. 2009b; 100:278–290. [PubMed: 20694052]
- You J, Chen G, Zhou Y. Block empirical likelihood for longitudinal partially linear regression models. *Canadian Journal of Statistics*. 2006; 34:79–96.

Figure 1.

True and estimated nonparametric function $\hat{\theta}(t)$ based on the 2000 simulations: (—True; -- the proposed method (Method 1); $\cdots$ ignoring the population-level information (Method 2))

Figure 2.

Empirical pointwise standard errors of the estimated nonparametric function $\widehat{\theta}(t)$ based on the 2000 simulations: (—the proposed method (Method 1); - - - ignoring the population-level information (Method 2))

Figure 3.

Estimated kernel estimate $\hat{\theta}(\text{age})$ when fitting the semiparametric model to the dementia study and its 95% confidence intervals (— $\hat{\theta}(\text{age})$; - - - - 95% confidence interval)

Table 1
Empirical percent relative biases, standard deviations and coverage probabilities for four different methods

n	ρ	α ₁	Para.	Method 1			Method 2			Method 3			Method 4		
				BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP
The true model is non-linear															
400	0.6	-5	β ₁	1.86	0.243	94.6	2.69	0.254	95.2	-5.43	0.289	90.2	-5.01	0.247	91.8
			β ₂	-0.72	0.259	95.0	-2.79	0.262	93.4	5.04	0.301	92.2	6.48	0.263	90.8
		-2	β ₁	-1.66	0.115	94.6	-1.67	0.115	94.4	-4.05	0.109	90.2	-4.04	0.109	89.8
			β ₂	1.35	0.115	94.8	4.30	0.115	94.6	8.48	0.107	86.4	8.52	0.107	86.6
		0.0	β ₁	1.82	0.234	94.6	3.82	0.234	94.3	-1.47	0.255	92.4	-1.36	0.241	93.4
			β ₂	-1.47	0.244	94.2	-6.57	0.242	94.0	0.81	0.261	92.0	1.01	0.250	92.6
	200	-5	β ₁	0.83	0.107	94.6	0.85	0.107	93.8	-1.96	0.100	94.2	-1.97	0.100	94.0
			β ₂	-1.25	0.108	94.2	-1.26	0.108	94.6	3.49	0.102	94.0	3.50	0.103	93.8
		-2	β ₁	0.46	0.312	94.8	1.65	0.320	94.8	-7.22	0.370	92.6	-7.30	0.312	91.6
			β ₂	-2.48	0.337	94.8	-2.44	0.340	94.3	7.37	0.391	90.4	7.57	0.339	91.2
		0.0	β ₁	-1.67	0.158	94.6	-1.64	0.158	94.4	-4.06	0.148	92.8	-4.13	0.148	93.2
			β ₂	2.74	0.174	94.4	2.67	0.174	94.4	9.07	0.160	89.2	9.13	0.160	89.0
The true model is linear															
400	0.6	-5	β ₁	1.65	0.247	94.8	2.33	0.258	94.8	-0.36	0.248	94.4	0.02	0.307	94.4
			β ₂	2.46	0.262	95.0	0.19	0.270	94.6	2.17	0.271	93.8	2.30	0.317	94.6
		-2	β ₁	-1.74	0.107	94.0	-1.71	0.108	94.8	-2.23	0.105	93.8	-2.21	0.105	93.6
			β ₂	1.55	0.114	94.4	1.50	0.114	94.4	2.30	0.112	94.2	5.24	0.112	94.2
		0.0	β ₁	1.41	0.239	94.6	1.69	0.241	95.8	0.88	0.257	94.2	1.25	0.272	94.2
			β ₂	-1.70	0.244	94.8	-1.04	0.246	95.8	-1.29	0.257	94.8	-1.66	0.269	94.8
	200	-5	β ₁	0.57	0.107	94.4	0.56	0.107	95.2	0.03	0.106	95.8	0.01	0.106	95.6
			β ₂												
		-2	β ₁												
			β ₂												
		0.0	β ₁												
			β ₂												

n	ρ	α_1	Para.	Method 1			Method 2			Method 3			Method 4		
				BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP
	β_2	-2		-0.78	0.113	95.0	-0.79	0.113	94.2	0.03	0.112	94.0	0.01	0.112	94.0

Table 2

Empirical percent relative biases, standard deviations and coverage probabilities for four different methods with a misspecified missing data model: sample size is 400

ρ	α_1	Para.	Method 1			Method 2			Method 3			Method 4		
			BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP	BIAS	SD	CP
0.6	-5	β_1	2.171	0.136	84.6	3.173	0.138	85.4	-5.157	0.127	92.6	-3.560	0.126	93.8
		β_2	-8.617	0.142	77.6	-8.101	0.144	81.4	8.375	0.130	90.4	3.832	0.128	94.0
	-2	β_1	-2.168	0.101	91.4	-2.188	0.103	91.4	-4.861	0.098	91.2	-4.785	0.098	91.2
		β_2	5.939	0.109	87.0	5.996	0.110	87.2	9.349	0.101	84.6	9.186	0.101	84.8
0.0	-5	β_1	0.754	0.133	84.8	0.728	0.133	84.0	-2.475	0.127	95.0	-2.436	0.127	95.0
		β_2	-1.923	0.145	83.0	-1.915	0.145	82.6	3.625	0.136	94.0	3.648	0.136	93.6
	-2	β_1	0.730	0.102	93.0	0.721	0.102	93.0	-1.835	0.098	94.8	-1.837	0.098	94.8
		β_2	-0.964	0.105	90.8	-0.968	0.105	91.0	3.603	0.102	94.4	3.613	0.102	94.6

Table 3

Analysis results for the dementia study data: estimates of the parametric parameters in the response model and missing data model parameters

Para.	Method 1			Method 2			Method 3			Method 4		
	EST	SE	P	EST	SE	P	EST	SE	P	EST	SE	P
Response model												
SEX	-0.097	0.039	0.012	-0.105	0.041	0.010	-0.103	0.056	0.069	-0.133	0.057	0.020
DEPD	0.507	0.044	<0.001	0.510	0.047	<0.001	0.271	0.044	<0.001	0.357	0.044	<0.001
fhdem	0.035	0.042	0.414	0.038	0.045	0.398	0.036	0.041	0.380	0.044	0.042	0.297
diabete	-0.225	0.057	<0.001	-0.223	0.060	<0.001	-0.183	0.075	0.015	-0.199	0.077	0.010
hypert	-0.147	0.040	<0.001	-0.148	0.042	<0.001	-0.120	0.054	0.026	-0.142	0.055	0.010
EDUC	0.002	0.002	0.301	0.002	0.002	0.283	0.001	0.003	0.765	0.000	0.003	0.986
Missing data model												
(Intercept)	-1.862	0.265	<0.001									
SEX	-0.523	0.048	<0.001									
DEPD	1.207	0.061	<0.001									
fhdem	0.012	0.050	0.808									
diabete	0.201	0.072	0.005									
hypert	0.002	0.048	0.962									
EDUC	-0.011	0.003	<0.001									
AGE	0.037	0.003	<0.001									
Prev. dementia	4.522	0.318	<0.001									