Title

A semivarying joint model for longitudinal binary and continuous outcomes

Journal

Canadian Journal of Statistics, 44(1)

ISSN

0319-5724

Authors

Kürüm, Esra
Hughes, John
Li, Runze

Publication Date

2016-03-01

DOI

10.1002/cjs.11273

Peer reviewed

HHS Public Access

Author manuscript

Can J Stat. Author manuscript; available in PMC 2017 March 01.

Published in final edited form as:

Can J Stat. 2016 March ; 44(1): 44–57. doi:10.1002/cjs.11273.

A semivarying joint model for longitudinal binary and continuous outcomes

Esra Kürüm[1,*], **John Hughes**[2], and **Runze Li**[3]

[1]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06520, USA

[2]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

[3]Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, USA

Abstract

Semivarying models extend varying coefficient models by allowing some regression coefficients to be constant with respect to the underlying covariate(s). In this paper we develop a semivarying joint modelling framework for estimating the time-varying association between two intensively measured longitudinal response: a continuous one and a binary one. To overcome the major challenge of jointly modelling these responses, namely, the lack of a natural multivariate distribution, we introduce a Gaussian latent variable underlying the binary response. Then we decompose the model into two components: a marginal model for the continuous response, and a conditional model for the binary response given the continuous response. We develop a two-stage estimation procedure and discuss the asymptotic normality of the resulting estimators. We assess the finite-sample performance of our procedure using a simulation study, and we illustrate our method by analyzing binary and continuous responses from the Women's Interagency HIV Study.

Key words and phrases

Generalized varying coefficient model; HIV; local linear regression; profile least squares

1. INTRODUCTION

Analysis of longitudinal data can be challenging due to intra-subject dependence. When there are multiple responses and the association between those responses is of interest, it is common to model the responses jointly. In applications of joint modelling of longitudinal responses, two challenges may be encountered: (1) the responses may be of different types (such as binary and continuous), in which case no natural multivariate distribution exists, and (2) the data may exhibit a dynamic pattern that cannot be revealed by ordinary models. These issues imply the need for a general statistical procedure for analyzing longitudinal binary and continuous outcomes, a procedure that permits the association between the

[*]Author to whom correspondence may be addressed: esrakurum@gmail.com.

responses to be time varying while accommodating two types of response–predictor relationships: time varying and time invariant. In this paper we propose such a procedure.

Various methods have been developed for modelling longitudinal binary and continuous responses jointly (Catalano and Ryan, 1992; Cox and Wermuth, 1992; Dunson, 2000; Fitzmaurice and Laird, 1995; Gueorguieva and Agresti, 2001; Kürüm et al., 2014; Liu et al., 2009; Regan and Catalano, 1999; Sammel et al., 1997). The chief difficulty in developing these methods is that there is no natural multivariate distribution for such outcomes. One solution to this problem is to introduce a latent variable underlying the binary response, and assume that the continuous response and the latent variable are jointly normally distributed. This joint distribution is then decomposed in one of two ways: (1) a marginal distribution for the continuous response and a conditional distribution for the binary response given the continuous response, or (2) a marginal distribution for the binary response and a conditional distribution for the continuous response given the binary response.

Another solution to the aforementioned problem is the joint mixed-effects model (Gueorguieva, 2001; Gueorguieva and Agresti, 2001). In this model a random effect is assumed for each response, and the responses are associated through a joint distribution for the random effects. One disadvantage of this approach is that maximum likelihood estimation is possible only when strong assumptions are made (Verbeke et al., 2010). For instance, Roy and Lin (2000) assumed that the random effects are perfectly correlated. Moreover, a mixed-effects model may be confounded (Hodges and Reich, 2010), which may inflate the variance of fixed-effects estimators and thereby prevent the discovery of important response–predictor relationships.

In a longitudinal study the relationship between a response and predictors, or the association between the continuous and binary responses, may vary over time, and ordinary models cannot capture these dynamic patterns. For this reason, unlike the aforementioned joint modelling techniques, Kürüm et al. (2014) proposed time-varying coefficient models (Brumback and Rice, 1998; Hoover et al., 1998) for modelling longitudinal binary and continuous responses jointly. Their method allows all parameters, including association parameters, to be time varying. These nonparametric models relax the restrictive assumptions of parametric models and are very useful in exploring the hidden structure in a data set. A nonparametric approach may, however, lack power when the sample is small, in which case a semivarying model may be more appropriate, especially if the practitioner has reason to believe that some response–predictor relationships are time invariant.

In this paper we introduce a new joint modelling approach for intensively measured longitudinal binary and continuous outcomes. The goals of our approach are to (1) efficiently estimate the time-varying partial association between the responses conditional on predictors of interest, and to (2) reveal both time-varying and time-invariant response–predictor relationships.

To achieve our goals, we employ semivarying coefficient models, which were studied by Fan and Huang (2005) for independent and identically distributed (iid) observations, and were extended to longitudinal data analysis by Fan et al. (2007). These models are extensions of

both partially linear models (Härdle et al., 2000) and time-varying coefficient models (Hastie and Tibshirani, 1993; Hoover et al., 1998). The major contribution of this work is that our proposed approach can explore the time-varying partial association between binary and continuous responses while allowing some of the response–predictor relationships to be time invariant, unlike the approach of Kürüm et al. (2014), which posits that all of these relationships are time-varying. To the best of our knowledge, semivarying coefficient models have not been applied for jointly modelling these types of responses in a longitudinal setting. This work will fill this gap in the literature.

We propose an estimation procedure for semivarying coefficient models for joint binary and continuous outcomes. We adopt the above mentioned latent variable approach and then factor the joint distribution into two components. This results in a two-stage estimation procedure. In the first stage we fit the marginal model of the continuous response by using estimation techniques for semivarying coefficient models. In the second stage we use generalized time-varying coefficient models (Cai et al., 2000) (for iid data) to fit the conditional model of the binary response given the continuous response. We use a simulation study to investigate the efficacy of our procedure.

The remainder of this paper is organized as follows. Section 2 introduces our joint model for longitudinal binary and continuous responses, and describes our two-stage estimation procedure. In this section, we also discuss the asymptotic behavior of our estimators. Section 3 presents our simulation study. Section 4 illustrates our proposed methodology by analyzing data from the Women's Interagency HIV Study. In Section 5 we make concluding remarks.

2. JOINT MODEL FOR BINARY AND CONTINUOUS RESPONSES

2.1. Model Specification

When the joint distribution of two continuous responses is of interest, it may be reasonable to assume that the responses are bivariate normal. There is no analogous joint distribution for the binary–continuous case, however. In order to overcome this challenge, we follow the well-known joint modelling approach described in Catalano and Ryan (1992). That is, we introduce a latent variable underlying the binary response, and assume that the continuous response and the latent variable are jointly normal. We then obtain the desired joint distribution using a two-component factorization: a marginal model for the continuous variable, and a conditional model for the binary variable given the continuous variable. The first component is readily available, and the second component is obtained using the normality of the latent variable along with the relationship defined between the latent variable and the binary response.

Before we describe our modelling scheme, let us introduce some notation. Suppose we have n independent subjects. For subject i, let $Q_i(t)$ and $W_i(t)$ denote the binary and continuous responses, respectively, measured at time points $t = t_{ij}$, where $j = 1, \ldots, n_i$ and n_i is the number of observations for subject i. We denote the latent variable as $Y_i(t)$. Let $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^{\mathrm{T}}$ and $\mathbf{z}_i(t) = (z_{i1}(t), \ldots, z_{iq}(t))^{\mathrm{T}}$ be the vectors of predictors for subject i. To simplify our presentation, we use the same set of predictors for both the continuous response

and the latent variable, but our method can handle different predictors for the two responses, which we demonstrate in the data application.

Now, consider the bivariate semivarying coefficient model

$$\begin{aligned} W_i(t) &= \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_w(t) + \mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_w + \varepsilon_{wi}(t) \\ Y_i(t) &= \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_y(t) + \mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_y + \varepsilon_{yi}(t), \end{aligned} \quad (1)$$

where $\boldsymbol{\beta}_\bullet = (\beta_{\bullet 1}, \ldots, \beta_{\bullet q})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_\bullet(t) = (\alpha_{\bullet 1}(t), \ldots, \alpha_{\bullet p}(t))^{\mathrm{T}}$ are the unknown regression coefficient vector and the nonparametric smooth baseline functions, respectively, and both $\varepsilon_{wi}(t)$ and $\varepsilon_{yi}(t)$ follow normal distributions with mean zero and time-varying variances $\sigma_w^2(t)$ and $\sigma_y^2(t)$, respectively. Let $\rho_w(\cdot, \cdot)$ and $\rho_y(\cdot, \cdot)$ be the correlations between two errors measured at different time points for $\varepsilon_{wi}(\cdot)$ and $\varepsilon_{yi}(\cdot)$, respectively. $\tau(t) = \mathrm{corr}\{\varepsilon_{wi}(t), \varepsilon_{yi}(t)\}$ is the partial correlation between $W_i(t)$ and $Y_i(t)$ given $\mathbf{x}_i$ and $\mathbf{z}_i$. Thus, we refer to $\tau(t)$ as the partial association between the binary and continuous responses. The primary goal of this paper is to develop an accurate estimator of this association.

In our modelling scheme, the relation between the binary variable and the latent variable is defined as: $Q_i(t) = 1$ if $Y_i(t) > 0$, and $Q_i(t) = 0$ if $Y_i(t) \le 0$. Since $Y_i(t)$ follows a normal distribution, the binary response $Q_i(t)$ follows the probit model

$$P\{Q_i(t)=1|\mathbf{x}_i(t), \mathbf{z}_i(t)\} = \Phi\left\{\frac{\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_y(t) + \mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_y}{\sigma_y(t)}\right\}, \quad (2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

To obtain the joint distribution, we use the two-component decomposition with a marginal model for the continuous variable $W_i(t)$, and a conditional model for $Q_i(t)$ given $W_i(t)$:

$$f_{q,w}\{q_i(t), w_i(t)\} = f_w\{w_i(t)\}\, f_{q|w}\{q_i(t)|w_i(t)\}.$$

The first component (the marginal model for the continuous response) was defined in (1). The second component (the conditional model for $Q_i(t)$ given $W_i(t)$) is derived using the conditional model for $Y_i(t)$ given $W_i(t)$. Normal theory shows that the conditional distribution $Y_i(t) \mid W_i(t)$ follows a Gaussian distribution, and the mean of this conditional distribution depends on the error from the marginal model of the continuous response. Specifically,

$$Y_i(t)|W_i(t) \sim \mathcal{N}\left[\mu_i(t),\, \sigma_y^2(t)\{1-\tau^2(t)\}\right], \quad (3)$$

where

$$\mu_i(t)=\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_y(t)+\mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_y+\frac{\sigma_y(t)}{\sigma_w(t)}\tau(t)\,\varepsilon_{wi}(t)$$

and

$$\varepsilon_{wi}(t)=W_i(t)=\{\mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_w(t)+\mathbf{z}_i^{\mathrm{T}}(t)\beta_w\}$$

is the error from the marginal model of the continuous response. Combining (3) with the relation defined between the latent variable and the binary response leads to the model

$$P\{Q_i(t)=1|W_i(t)\}=\Phi\left[\frac{\mu_i(t)}{\sqrt{\sigma_y^2(t)\{1-\tau^2(t)\}}}\right]. \quad (4)$$

Note that not all of the parameters in model (4) are estimable. For instance, it is not possible to estimate $\boldsymbol{\alpha}_y(t)$ and $\sigma_y(t)$ separately, but the ratio $\boldsymbol{\alpha}_y(t)/\sqrt{\sigma_y^2(t)\{1-\tau^2(t)\}}$ is identifiable. Moreover, we might expect practitioners be more interested in the relationship between the predictors and the binary response. Thus we reparameterize the probit model in (4) to arrive at a more parsimonious and fully estimable form:

$$P\{Q_i(t)=1|W_i(t)\}=\Phi\{\mathbf{x}_i^{*\mathrm{T}}(t)\boldsymbol{\gamma}(t)+\gamma_{p+q+1}(t)\,\varepsilon_{wi}(t)\}, \quad (5)$$

where $\mathbf{x}_i^{*\mathrm{T}}(t)=(\mathbf{x}_i^{\mathrm{T}}(t),\mathbf{z}_i^{\mathrm{T}}(t))^{\mathrm{T}}$ and $\boldsymbol{\gamma}(t)=(\gamma_1(t),\ldots,\gamma_{p+q}(t))^{\mathrm{T}}$. The conditional form above shows that the continuous response is linked with the binary response in a probit regression model that includes the error from the marginal model as a covariate.

Model (5) and the definition of $\mu_i(t)$ lead to

$$\gamma_{p+q+1}(t)=\frac{1}{\sigma_w(t)}\cdot\frac{\tau(t)}{\sqrt{1-\tau^2(t)}}$$

and, hence,

$$\tau(t)=\frac{b(t)}{\sqrt{1+b^2(t)}}, \quad (6)$$

where $b(t)=\gamma_{p+q+1}(t)\sigma_w(t)$. According to (6), the partial association $\tau(t)$ depends on the regression coefficient of the error term from the marginal model and on the variance of the

continuous response. The regression coefficient $\gamma_{p+q+1}(t)$ and $\tau(t)$ share the same sign and are positively correlated.

2.2. Estimation Procedure

We propose a two-stage estimation procedure. Before presenting the details of our procedure, we give a brief sketch of both stages. In the first stage we fit a semivarying coefficient model (Fan and Huang, 2005; Fan et al., 2007) to the continuous response. At this stage we employ the profile least squares approach proposed by Fan et al. (2007) to obtain efficient estimators of the regression coefficients $\alpha_w(t)$ and β_w. In the second stage we use the residuals from the first stage and the predictors for the binary response, and fit a generalized varying coefficient model for the binary response given the continuous response. At this stage we obtain the components necessary to compute the estimate of $\tau(t)$.

Now we turn to the details for the first stage. Fan et al. (2007) suggest using the following formula to estimate the nonparametric smooth baseline functions in (1):

$$\hat{\boldsymbol{\alpha}}_w(t)=(\mathbf{I}_p,\mathbf{0}_p)(\boldsymbol{\Lambda}^{\mathrm{T}}\kappa\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\kappa\mathbf{W}^{*}, \quad (7)$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\mathbf{0}_p$ is the $p \times p$ matrix of zeros, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_n)^{\mathrm{T}}$, $\boldsymbol{\Lambda}_i = ((1, t_{i1} - t) \otimes \mathbf{x}_{i1}, \ldots, (1, t_{in_i} - t) \otimes \mathbf{x}_{in_i})$, and κ is an $N \times N$ diagonal matrix with the kernel weights along its diagonal.

Substituting $\hat{\alpha}_w(t)$ in (1) and using weighted least squares yields

$$\hat{\boldsymbol{\beta}}_w=\{\mathbf{z}^{\mathrm{T}}(\mathbf{I}_N-\mathbf{S})^{\mathrm{T}}\mathbf{R}(\mathbf{I}_N-\mathbf{S})\mathbf{z}\}^{-1}\mathbf{z}^{\mathrm{T}}(\mathbf{I}_N-\mathbf{S})^{\mathrm{T}}\mathbf{R}(\mathbf{I}_N-\mathbf{S})\mathbf{W},$$

where $\mathbf{z}=(\mathbf{z}_1^{\mathrm{T}},\ldots,\mathbf{z}_n^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{W}=(\mathbf{W}_1^{\mathrm{T}},\ldots,\mathbf{W}_n^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{W}_i = (W_i(t_{i1}), \ldots, W_i(t_{in_i}))^{\mathrm{T}}$, $\mathbf{R}$ is the working covariance matrix, and $\mathbf{S}$ is the smoothing matrix of the local linear smoother. Misspecification of the working covariance matrix affects only the efficiency, not the consistency, of this estimator, whereas the local linear estimator (7) is not significantly affected by the covariance structure since the data are localized in time (Fan et al., 2007). We demonstrate this result by using various covariance structures in our simulation study.

After we fit a semivarying coefficient model to the continuous response and obtain the residuals from this fit, we move to the second stage. In the second stage we fit a generalized time-varying coefficient model for the conditional model (5). Cai et al. (2000) introduced generalized varying coefficient models for independent and identically distributed data. We adapt these models to a longitudinal setting.

We start by locally approximating the functions in a neighbourhood of a fixed point t_0 via the Taylor expansion:

$$\gamma_r(t) \approx \gamma_r(t_0)+\gamma_r'(t_0)(t-t_0) \equiv a_r^*+b_r^*(t-t_0), \quad (8)$$

for $r = 1, \ldots, p + q + 1$. Let $\mathbf{a}^*=(a_1^*, \ldots, a_{p+q+1}^*)^{\mathrm{T}}$ and $\mathbf{b}^*=(b_1^*, \ldots, b_{p+q+1}^*)^{\mathrm{T}}$. For subject i, let $\mathbf{x}_i^*(t)=(\mathbf{x}_i^{\mathrm{T}}(t), \mathbf{z}_i^{\mathrm{T}}(t), e_i(t))^{\mathrm{T}}$. We maximize the local likelihood

$$\ell_n(\mathbf{a}^*, \mathbf{b}^*)=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\ell\left(g^{-1}\left[\sum_{r=0}^{p+q+1}\{a_r^*+b_r^*(t-t_0)\}\, x_{ir}^*(t)\right], Q_i(t)\right) K_{h_2}(t-t_0), \quad (9)$$

where $g(\cdot)$ is a link function, which is probit for our procedure, and h_2 is the bandwidth for the second stage. We use an iterative regression algorithm to find the solution that satisfies $\ell'(\mathbf{a}^*, \mathbf{b}^*) = 0$, and the estimators are given by $\hat{\mathbf{a}}^* = \hat{\boldsymbol{\gamma}}(t_0) = (\hat{\gamma}_1(t_0), \ldots, \hat{\gamma}_{p+q+1}(t_0))^{\mathrm{T}}$. Details of this algorithm are presented in the supplementary material.

It is necessary to derive pointwise confidence intervals for the nonparametric components in both stages of the estimation procedure, and to do so we need estimates of the asymptotic covariance matrices. We encourage the reader to refer to the supplementary material for details on these estimators. The sampling properties of the estimators obtained in both stages are also discussed in the supplementary material.

Recall that our chief goal is to estimate the partial association $\tau(t_0)$ between the continuous and binary responses. According to (6), to obtain $\hat{\tau}(t_0)$ we need to estimate $\gamma_{p+q+1}(t_0)$ and $\sigma_w^2(t_0)$. The estimator of a_{p+q+1}^* gives us $\hat{\gamma}_{p+q+1}(t_0)$. To estimate $\sigma_w^2(t_0)$ we propose using the kernel estimator

$$\hat{\sigma}_w^2(t_0)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} e_i^2(t_{ij})K_{h_1}(t_{ij}-t_0)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i} K_{h_1}(t_{ij}-t_0)}. \quad (10)$$

Plugging $\hat{\sigma}_w(t_0)$ and $\hat{\gamma}_{p+q+1}(t_0)$ into (6) gives the estimator for $\tau(t_0)$. Based on the relationship between $\gamma_{p+q+1}(t_0)$ and $\tau(t_0)$ defined in (6), the pointwise asymptotic confidence band for $\gamma_{p+q+1}(t_0)$ gives us information regarding the significance of the partial association $\tau(t_0)$. Another way of determining the significance of the partial association is to use a nonparametric bootstrapping procedure. As we demonstrate in the data application, the bootstrap and asymptotic confidence bands agree closely.

For methods based on kernel smoothing, selecting a suitable bandwidth and kernel function are important. These issues are addressed in detail in the supplementary material.

3. SIMULATED APPLICATION

In this section we demonstrate the performance of our proposed procedure via a Monte Carlo simulation study that mimics the data application presented in Section 4. For the simulation study we used the $K_{0.1}$ bimodal kernel and a set of equidistant grid points $\{t_k, k = 1, \ldots, n_{\text{grid}}\}$ between 0 and 1 with $n_{\text{grid}} = 200$. We simulated 500 intensive longitudinal data sets. For the ith subject, the number of observations n_i was randomly selected according to the discrete uniform distribution on [1, 8], and the measurement times $T_i = (t_{i1}, \ldots, t_{in_i})$ were drawn from the standard uniform distribution. We used a sample size of $n = 300$.

Let $\mathbf{x}_{wi}(t) = (1, x_{w1i}(t))^{\mathrm{T}}$, $\mathbf{x}_{yi}(t) = (1, x_{y1i}(t))^{\mathrm{T}}$, $\mathbf{z}_{wi}(t) = (z_{w1i}(t), z_{w2i}(t), z_{w3i}(t))^{\mathrm{T}}$, and $z_{yi}(t) = z_{y1i}(t)$. We then generated the continuous and latent variables from the following models:

$$\begin{aligned} W_i(t) &= \mathbf{x}_{wi}^{\mathrm{T}}(t)\boldsymbol{\alpha}_w(t) + \mathbf{z}_{wi}^{\mathrm{T}}(t)\boldsymbol{\beta}_w + \varepsilon_{wi}(t), \\ Y_i(t) &= \mathbf{x}_{yi}^{\mathrm{T}}(t)\boldsymbol{\alpha}_y(t) + z_{yi}(t)\beta_y + \varepsilon_{yi}(t), \end{aligned} \tag{11}$$

where $\boldsymbol{\alpha}_w(t) = (\alpha_{w0}(t), \alpha_{w1}(t))^{\mathrm{T}} = (\sin(2\pi t), \cos(2\pi t))^{\mathrm{T}}$, $\boldsymbol{\beta}_w = (\beta_{w1}, \beta_{w2}, \beta_{w3})^{\mathrm{T}} = (0.3, 0.15, -0.10)^{\mathrm{T}}$, $\boldsymbol{\alpha}_y(t) = (\alpha_{y0}(t), \alpha_{y1}(t))^{\mathrm{T}} = (-0.2 \sin(2\pi t), 1 + \cos(2\pi t))^{\mathrm{T}}$, $\beta_y = -0.05$, and $i = 1, \ldots, 300$. We simulated the predictors from the standard Gaussian distribution. Both $\varepsilon_{wi}(t)$ and $\varepsilon_{yi}(t)$ follow Gaussian distributions with mean zero and time-varying variance $0.4 + 0.4 \sin^2(2\pi t)$. The correlations between two error terms measured at different time points are $\rho_w(t_1, t_2) = 2^{-7|t_1 - t_2|}$ and $\rho_y(t_1, t_2) = 5^{-7|t_1 - t_2|}$ for $\varepsilon_{wi}(\cdot)$ and $\varepsilon_{yi}(\cdot)$, respectively. The association between the binary and continuous responses measured at time t is $\tau(t) = 0.2 + 0.15 \sin(2\pi t)$. The primary aim of our study was to demonstrate that we can accurately estimate this association.

In Section 2.1 the binary variable was defined as $Q_i(t_{ij}) = 1$ if $Y_i(t_{ij}) > 0$, and $Q_i(t_{ij}) = 0$ if $Y_i(t_{ij}) \leq 0$. However, it is of interest to show that decreasing the percentage of successes in the binary response does not decrease the efficacy of our procedure. Hence, the relation between the latent variable and the binary variable was defined as $Q_i(t_{ij}) = 1$ if $Y_i(t_{ij}) > 0.25$, and $Q_i(t_{ij}) = 0$ if $Y_i(t_{ij}) \leq 0.25$. Therefore, each of our 500 simulated data sets had approximately 40% failure.

In the first stage we fit a semivarying coefficient model to the marginal model of the continuous response. The estimation procedure requires that we choose a covariance structure. According to Fan et al. (2007), the performance of estimators for the parametric and nonparametric components should be similar no matter the chosen structure. To show that this result holds for our approach, we compared results for three covariance structures: the identity matrix, an ARMA(1, 1) structure, and the true within-subject covariance structure of the continuous response. The results were comparable, and so we present the results for the true covariance structure only.

We generated several pilot data sets and used the leave-one-out cross-validation bandwidth selector to obtain the optimal bandwidth. Figure 1 shows the results for $h_1 = 0.075$: the bandwidth that minimized the cross-validation score. Our estimation procedure for the first stage performed well with respect to bias, as the biases are typically near zero. We used a

smaller bandwidth for variance estimation in order to obtain more accurate estimates. Specifically, we used h_0 in $O(n^{-1/4})$ since the asymptotically optimal bandwidth is in $O(n^{-1/5})$. We see from the plots in Figure 1 that this bandwidth yielded accurate confidence intervals. We estimated parametric components $\hat{\beta}_{wj}$ (j = 1, 2, 3) with small biases and mean squared errors of (0.003, 0.001, 0.001) and (0.002, 0.003, 0.002), respectively.

In the second stage we fit a generalized time-varying coefficient model to the conditional model

$$P\{Q_i(t)=1|W_i(t)\}=\Phi\{\mathbf{x}_i^{*\mathrm{T}}(t)\boldsymbol{\gamma}(t)+\gamma_4(t)e_{wi}(t)\},$$

where $\boldsymbol{\gamma}(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))^{\mathrm{T}}$ is the vector of varying coefficient functions, and $\mathbf{x}_i^{*\mathrm{T}}(t)=(\mathbf{x}_{yi}^{\mathrm{T}}(t), z_{yi}(t))^{\mathrm{T}}$ and $e_{wi}(t)=W_i(t)-\{\mathbf{x}_{wi}^{\mathrm{T}}(t)\hat{\boldsymbol{\alpha}}_w(t)+\mathbf{z}_{wi}^{\mathrm{T}}(t)\hat{\boldsymbol{\beta}}_w\}$ are the predictors and the residual, respectively, from the first stage, for subject i.

We once again used a pilot study to choose the optimal bandwidth. At this stage we obtain $\hat{\gamma}_k^*(t)(k=1,\ldots,4)$ at the optimal bandwidth (h_2 = 0.15), and the kernel estimate of $\sigma_w^2(t)$ at the optimal bandwidth for the first stage (h_1 = 0.075). We then estimate $\tau(t)$ using (6). Figure 2 shows the median estimated time-varying partial association based on 500 Monte Carlo simulation runs, along with the 2.5 and 97.5 percentiles based on 500 bootstrap samples. Judging from this plot, the median estimated time-varying association is close to the true association.

4. APPLICATION TO DATA FROM THE WOMEN'S INTERAGENCY HIV STUDY

We now illustrate our proposed joint modelling methodology via an analysis of data from the Women's Interagency HIV Study (WIHS). The data are for 372 women recruited from HIV testing sites in Chicago, San Francisco, Los Angeles, New York City (Bronx and Brooklyn), and Washington, DC between 1994 and 1995. Participants were scheduled to have a semiannual interview at a WIHS site. During this interview, participants received physical and oral examinations, gave blood, urine, and gynecological specimens, and also answered a series of questions about their daily activities such as sexual behaviors, tobacco and alcohol use. We restricted our analysis to 292 HIV positive participants aged 25–55. 26%, 45%, and 12% of these subjects self-identified as Latina or Hispanic, African-American non-Hispanic origin, and white non-Hispanic origin, respectively. Among the participants, 66% were smokers, and by the end of the study 8.3% of the smokers quit smoking whereas 8.1% of the non-smokers started smoking. Our data set contains follow-up information on the participants until 2006. Since many participants missed some of their scheduled visits, the number of measurements and measurement times varies from subject to subject. The number of observations for each participant varies from one to eight.

Ferson et al. (1979), Galai et al. (1997), Halonen et al. (1982) and Hughes et al. (1985) demonstrated that cigarette smoking has effects on the immune system, but it is not yet clear whether any of these effects influence the progression of HIV to AIDS. Galai et al. (1997)

analyzed data from the Multicenter AIDS Cohort Study of homosexual men in order to investigate the effect of cigarette smoking on development of AIDS. They applied Kaplan–Meier analysis and multivariate Cox regression models, and concluded that smoking was not significantly associated with progression to AIDS. Likewise, Burns et al. (1996) studied the association between cigarette smoking and HIV progression on a cohort of 3,221 HIV-seropositive men and women enrolled in the Terry Beirn Community Programs for Clinical Research on AIDS. Using proportional hazards regression analysis, they found no association between smoking and the overall risk of HIV progression or death. On the other hand, Nieman et al. (1993) showed that in a case series of 84 individuals, smokers progressed to AIDS more rapidly than nonsmokers. They employed life tables and compared median time to develop AIDS for smokers and nonsmokers. Our main interest is to study the relationship between HIV progression (measured by CD4 cell percentage) and smoking status for the women who participated in the WIHS.

There are two important differences between our approach and the above mentioned methods. First, some of the previous analyses excluded subjects who changed their smoking behavior during the study. The drawback of this exclusion is that in a longitudinal study we expect behaviors to change over time, and the inability to take these changes into account may result in biased results. Second, we do not apply survival methods, since our data are not censored. In addition, we are interested in investigating the relationship between CD4 cell percentage and smoking status throughout the study instead of defining a lifetime for a subject that ends when the subject progresses to AIDS.

Based on some preliminary analysis and findings in the HIV literature (Zeger and Diggle, 1994; Obirikorang and Yeboah, 2009), we chose a set of predictors for each response. The predictors for CD4 cell percentage (the continuous response) were: baseline CD4 cell percentage (measured at the first visit), number of sexual partners, hematocrit value (the volume percentage of red cells in the blood), mean corpuscular volume (a measure of average red blood cell size), platelet count, and Center for Epidemiologic Studies Depression (CESD) scale score. For smoking status (the binary response), CESD scale score and race were used as predictors. All predictors except race are continuous variables, and they were centered. Note that the race variable initially had five levels: African American, white, Asian/Pacific Islander, native American/Alaskan native, and other. Since our data set had only two participants in each of the Asian/Pacific Islander and native American/Alaskan native categories, we recategorized race into three levels: African American, white, and other.

We began by determining whether a fully time-varying or a semivarying approach is appropriate for these data. To do so we fit the appropriate time-varying coefficient model to each of the continuous and binary responses. Since some of the effects appeared to be constant over time, we decided to proceed with a semivarying analysis. For the continuous response, we observed that the intercept and baseline CD4 cell percentage have time-varying effects, while the remaining effects are time invariant. For the binary response, CESD and the first dummy variable for race (RACE1), which is equal to 1 if a subject is African American, were time varying, while the second dummy variable for race (RACE2), which is equal to 1 if a subject is white, was time invariant.

In our analysis the $K_{0,1}$ bimodal kernel function was used in both stages. We applied the leave-one-out cross-validation method, and chose $h_1 = 29.5$ and $h_2 = 36$ as the bandwidths for the first and second stages, respectively.

In the first stage of our estimation procedure, we fit the following semivarying coefficient model to CD4 cell percentage:

$$W_i(t) = \mathbf{x}_i^{\mathrm{T}}(t)\boldsymbol{\alpha}_w(t) + \mathbf{z}_i^{\mathrm{T}}(t)\boldsymbol{\beta}_w + \varepsilon_{wi}(t),$$

where $\boldsymbol{\alpha}_w(t) = (\alpha_{w0}(t), \alpha_{w1}(t))^{\mathrm{T}}$, $\boldsymbol{\beta}_w = (\beta_{w1}, \beta_{w2}, \beta_{w3}, \beta_{w4}, \beta_{w5})^{\mathrm{T}}$, $\mathbf{x}_i(t) = (1, x_{i1}(t))^{\mathrm{T}}$ with $x_{i1}(t)$ as the baseline CD4 cell percentage of subject i at the first visit, and $\mathbf{z}_i(t) = (z_{i1}(t), z_{i2}(t), z_{i3}(t), z_{i4}(t), z_{i5}(t))^{\mathrm{T}}$ with

$z_{i1}(t)$: the number of sexual partners of subject i at time t,

$z_{i2}(t)$: the hematocrit value of subject i at time t,

$z_{i3}(t)$: the mean corpuscular volume of subject i at time t,

$z_{i4}(t)$: the platelet count of subject i at time t,

$z_{i5}(t)$: the CESD scale score of subject i at time t,

and $t = t_{ij}$ the age of subject i at visit j.

The estimated time-varying functions in the first stage are depicted in Figure 3. The plot in panel (a) shows that the intercept function increases as age increases, and the confidence band suggests that the intercept function is time varying. From the plot in panel (b), we observe that the effect for baseline CD4 is time varying and decreases with age. Furthermore, the confidence intervals suggest that the effect is always significant and positive for ages between 25 and 55.

We next estimate the parametric component $\boldsymbol{\beta}_w$. Here we decided to use an ARMA(1, 1) correlation structure. Note that, as we would expect, using a working independence covariance matrix yielded similar results. The resulting estimates along with their corresponding 95% confidence intervals are displayed in Table 1. We see that all of the predictors except number of sexual partners are significantly associated with CD4 cell percentage for the WIHS data. The continuous response (CD4 cell percentage) has a positive relationship with the volume percentage of red cells in the blood (HCT), average red blood cell size (MCV), and the platelet count, and a negative relationship with CESD scale score.

After fitting the marginal model for the continuous response and obtaining the residuals for this fit, in the second stage of the estimation procedure we fit the conditional model for the binary response given the continuous response:

$$P\{Q_i(t) = 1 | W_i(t)\} = \Phi\{\mathbf{x}_i^{*\mathrm{T}}(t)\boldsymbol{\gamma}(t) + \gamma_5(t)e_{wi}(t)\},$$

where $Q_i(t)$ is the smoking status of subject i at time t, $e_{wi}(t)$ is the residual from the marginal fit of the continuous response, $\boldsymbol{\gamma}(t) = (\gamma_1(t), \ldots, \gamma_4(t))^{\mathrm{T}}$ is the vector of regression coefficients, and $\mathbf{x}_i^{*\mathrm{T}}(t)=(1, x_{2i}(t), x_{3i}(t), z_{6i}(t))^{\mathrm{T}}$ with

$x_{i2}(t)$: the CESD scale score of subject i at time t,

$x_{i3}(t)$: the first dummy variable for race ($x_{i2}(t) = 1$ if subject i is African American),

$z_{i6}(t)$: the second dummy variable for race ($z_{i6}(t) = 1$ if subject i is white).

As mentioned in Section 2.1, $\tau(t)=b(t)/\sqrt{1+b^2(t)}$ with $b(t) = \gamma_{p+q+1}(t)\sigma_w(t)$, where $\sigma_w^2(t)$ is the variance of CD4 cell percentage at time t, and $\tau(t)$ is the partial association between CD4 cell percentage and smoking status at time t conditional on a set of predictors. We estimate $\sigma_w(t)$ using kernel estimator (10) with bandwidth $h = 29.5$. To obtain a pointwise confidence band for $\tau(t)$, we generated 500 bootstrap samples by resampling from independent subjects.

Figure 4 (a) presents the estimated partial association $\hat{\tau}(t)$ along with 2.5 and 97.5 percentiles of the bootstrap samples. According to Figure 4 (a) we cannot conclude that the partial association between CD4 cell percentage and smoking status is time varying. However, we can conclude that the partial association is significant and negative for women between the ages of 29 and 46, i.e., for women enrolled in the WIHS, decreased CD4 cell percentage is partially associated with smoking. Based on the relationship between $\tau(t)$ and $\gamma_5(t)$, we can also investigate the significance of the partial association using the estimated confidence intervals for $\gamma_5(t)$. Figure 4 (b) depicts the estimated regression coefficient $\hat{\gamma}_5(t)$ along with its confidence band. This plot shows a slightly wider significance region for the association: the association is significant for women aged between 26 and 52, approximately.

Smoking is associated with pulmonary complications, decreased adherence to highly active antiretroviral therapy (Feldman et al., 2006), and increased incidence of opportunistic infections (Arcavi and Benowitz, 2004; Crothers et al., 2005; Kohli et al., 2006) for HIV-positive patients. Therefore, the results of our study and others suggest that successful smoking cessation programs are necessary for HIV patients in order to enhance the quality of life and make disease progression more manageable. A detailed review of existing smoking cessation techniques for HIV patients can be found in Niaura et al. (2012) along with ways to improve current research studies so that more effective cessation programs can be designed.

5. DISCUSSION

In this article we developed a new joint modelling approach for longitudinal binary and a continuous responses. In this approach the continuous response and the latent variable underlying the binary response are assumed to follow a semivarying coefficient model. We also proposed a two-stage estimation procedure based on local linear regression, and discussed the asymptotic normality of the resulting estimators in both stages. We

demonstrated that our procedure performs well for estimating the time-varying partial association between longitudinal binary and continuous responses. We also applied our methodology to a bivariate response (CD4 cell percentage and smoking status) from the Women's Interagency HIV Study. We concluded that there is a significant negative partial association between CD4 cell percentage and smoking status for women aged 29–46.

It might seem restrictive that our modelling scheme only considers a semivarying coefficient model for the continuous response but not for the binary response. However, in order to have a semivarying model for the binary response in (2), one has to assume that the coefficient of $\mathbf{z}_1(t)$ is proportional to the standard deviation function $\sigma_y(t)$. Such an assumption seems to be unnatural. Thus, it is more natural to consider a varying-coefficient probit model for the binary variable. Note that, in practice, an association may exist between the binary and continuous outcomes measured at different time points. Ignoring this dependence does not affect the asymptotic behavior of the estimators (Lin and Carroll, 2001; Kürüm et al., 2014).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Li's research was supported by NIH grants P50-DA10075, P50 DA039838, P50 DA036107 and R01 CA168676 and National Science Foundation grant DMS1512422. The content is solely the responsibility of the authors and does not necessarily represent the official views of these organizations. The authors are grateful to the Editor, the Associate Editor and two anonymous referees for providing valuable comments that significantly improved the paper.

BIBLIOGRAPHY

Arcavi L, Benowitz NL. Cigarette smoking and infection. Archives of Internal Medicine. 2004; 164(20):2206–2216. [PubMed: 15534156]

Brumback BA, Rice JA. Smoothing spline models for the analysis of nested and crossed samples of curves. Journal of American Statistical Association. 1998; 93(443):961–976.

Burns DN, Hillman D, Neaton JD, Sherer R, Mitchell T, Capps L, Vallier WG, Thurnherr MD. for the Terry Beirn Community Programs for Clinical Research on AIDS, FMG. Acquired Immune Deficiency Syndromes. 1996; 13(4):374–383.

Cai Z, Fan J, Li R. Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association. 2000; 95(451):888–902.

Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. Journal of the American Statistical Association. 1992; 87(419):651–658.

Cleveland, W.; Grosse, E.; Shyu, W. Local regression models. Wadsworth & Brooks/Cole; Pacific Grove, CA: 1992.

Cox DR, Wermuth N. Response models for mixed binary and quantitative variables. Biometrika. 1992; 79(3):441–461.

Crothers K, Griffith TA, McGinnis KA, Rodriguez-Barradas MC, Leaf DA, Weissman S, Gibert CL, Butt AA, Justice AC. The impact of cigarette smoking on mortality, quality of life, and comorbid illness among HIV-positive veterans. Journal of General Internal Medicine. 2005; 20(12):1142–1145. [PubMed: 16423106]

De Brabanter K, De Brabanter J, Suykens J, Moor BD. Kernel regression in the presence of correlated errors. Journal of Machine Learning Research. 2011; 12:1955–1976.

Dunson DB. Bayesian latent variable models for clustered mixed outcomes. Journal of the Royal Statistical Society. Series B. 2000; 62(2):355–366.

Fan J. Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics. 1993; 21(1):196–216.

Fan, J.; Gijbels, I. Local Polynomial Modelling and Its Applications. Chapman and Hall; London: 1996.

Fan J, Huang T. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli. 2005; 11(6):1031–1057.

Fan J, Huang T, Li R. Analysis of longitudinal data with semiparametric estimation of covariance function. Journal of American Statistical Association. 2007; 102(478):632–641.

Feldman JG, Minkoff H, Schneider MF, Gange SJ, Cohen M, Watts DH, Gandhi M, Mocharnuk RS, Anastos K. Association of cigarette smoking with HIV prognosis among women in the HAART era: A report from the Women's Interagency HIV Study. American Journal of Public Health. 2006; 96(6):1060–1065. [PubMed: 16670229]

Ferson M, Edwards A, Lind A, Milton GW, Hersey P. Low natural killer-cell activity and immunoglobulin levels associated with smoking in human subjects. International Journal of Cancer. 1979; 23(5):603–609. [PubMed: 457307]

Fitzmaurice GM, Laird NM. Regression models for a bivariate discrete and continuous outcome with clustering. Journal of the American Statistical Association. 1995; 90(431):845–852.

Galai N, Park LP, Wesch J, Visscher B, Riddler S, Margolick JB. Effect of smoking on the clinical progression of HIV-1 infection. Journal of Acquired Immune Deficiency Syndromes. 1997; 14(5): 451–8.

Gueorguieva R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. Statistical Modelling: An International Journal. 2001; 1(3): 177–193.

Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. Journal of the American Statistical Association. 2001; 96(455):1102–1112.

Halonen M, Barbee RA, Lebowitz MD, Burrows B. An epidemiologic study of the interrelationships of total serum immunoglobulin e, allergy skin-test reactivity, and eosinophilia. Journal of Allergy and Clinical Immunology. 1982; 69(2):221–228. [PubMed: 7056953]

Härdle, W.; Liang, H.; Gao, J. Partially Linear Models. New York: Springer-Verlag; 2000.

Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society. Series B. 1993; 55(4):757–796.

Hodges J, Reich B. Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician. 2010; 64(4):325–334.

Hoover DR, Rice JA, Wu CO, Yang LP. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika. 1998; 85(4):809–822.

Hughes DA, Haslam PL, Townsend PJ, Turner-Warwick M. Numerical and functional alterations in circulatory lymphocytes in cigarette smokers. Clinical & Experimental Immunology. 1985; 61(2): 459–466. [PubMed: 2931227]

Kohli R, Lo Y, Homel P, Flanigan TP, Gardner LI, Howard AA, Rompalo AM, Moskaleva G, Schuman P, Schoenbaum EE. Bacterial pneumonia, HIV therapy, and disease progression among HIV-infected women in the HIV epidemiologic research (her) study. Clinical Infectious Diseases. 2006; 43(1):90–98. [PubMed: 16758423]

Kürüm E, Li R, Shiffman S, Yao W. Time-varying coefficient models for joint modeling binary and continuous outcomes in longitudinal data. Statistica Sinica. in press.

Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimationg equations. Journal of American Statistical Association. 2001; 96(455):1045–1056.

Liu X, Daniels M, Marcus B. Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. Journal of the American Statistical Association. 2009; 104(486):429–438. [PubMed: 20161053]

Niaura R, Chander G, Hutton H, Stanton C. Interventions to address chronic disease and HIV: Strategies to promote smoking cessation among HIV-infected individuals. Current HIV/AIDS Reports. 2012; 9(4):375–384. [PubMed: 22972495]

Nieman RB, Fleming J, Coker RJ, William Harris JR, Mitchell DM. The effect of cigarette smoking on the development of AIDS in HIV-1-seropositive individuals. AIDS. 1993; 7(5):705–710. [PubMed: 8318178]

Obirikorang C, Yeboah FA. Blood haemoglobin measurement as a predictive indicator for the progression of HIV/AIDS in resource-limited setting. Journal of Biomedical Science. 2009; 16(1): 102. [PubMed: 19922646]

Regan MM, Catalano PJ. Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. Biometrics. 1999; 55(3):760–768. [PubMed: 11315004]

Roy J, Lin X. Latent variable models for longitudinal data with multiple continuous outcomes. Biometrics. 2000; 56(4):1047–1054. [PubMed: 11129460]

Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. Journal of the Royal Statistical Society: Series B. 1997; 59(3):667–678.

Severini TA, Staniswalis JG. Quasi-likelihood estimation in semiparametric models. Journal of American Statistical Association. 1994; 89(426)

Song, PX-K. Correlated Data Analysis: Modeling, Analytics, and Applications. Springer; New York: 2007.

Verbeke, G.; Molenberghs, G.; Rizopoulos, D. Random effects models for longitudinal data. In: van Montfort, K.; Oud, J.; Satorra, A., editors. Longitudinal Research with Latent Variables. Springer-Verlag; Berlin: 2010. p. 37-96.

Zeger SL, Diggle P. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics. 1994; 50(3):689–699. [PubMed: 7981395]

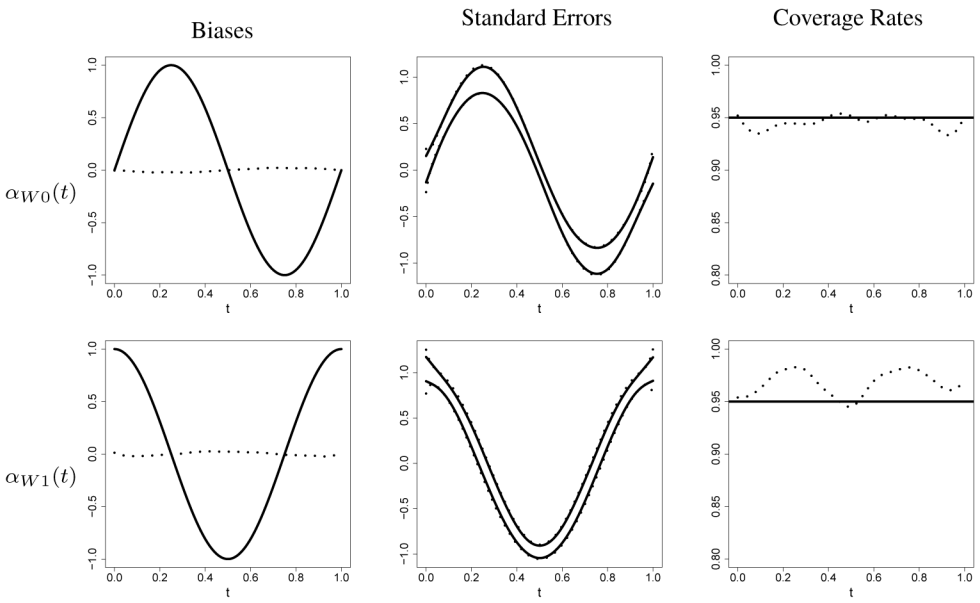

Figure 1.

Results from the first stage of our simulation study. Each row shows three plots for a given time-varying parameter. The first plot shows the true function (solid) and the empirical bias of our estimator (dotted). The second plot shows the empirical pointwise 95% confidence band (solid) and the mean theoretical pointwise 95% confidence band (dotted). The third plot shows the desired coverage rate (solid) and the empirical pointwise coverage rates (dotted).

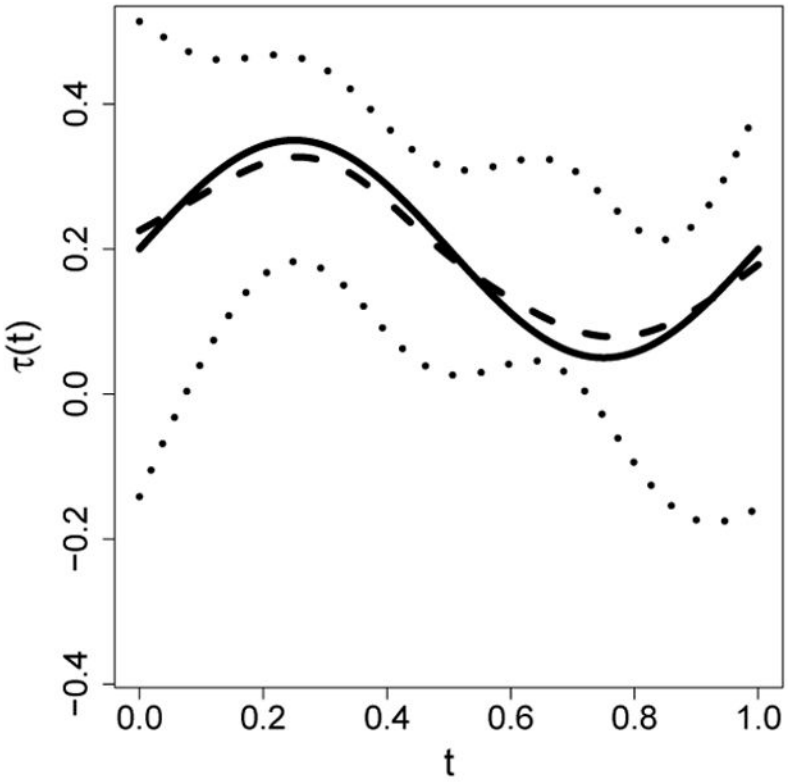

Figure 2.
Median estimated time-varying partial association (dashed) overlaying the true function (solid) along with 2.5 and 97.5 percentiles based on 500 bootstrap samples (dotted).

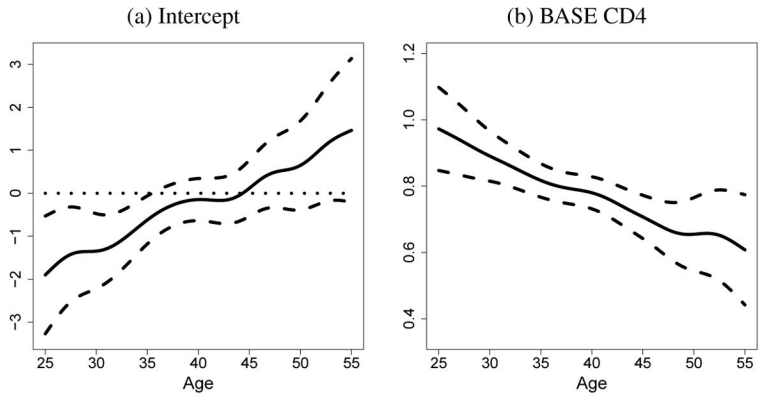

Figure 3.

The results of our data analysis for the continuous response, CD4 cell percentage. For each panel, the solid curve shows the estimate, the dashed curves show the estimated 95% pointwise confidence band, and the dotted line marks zero.

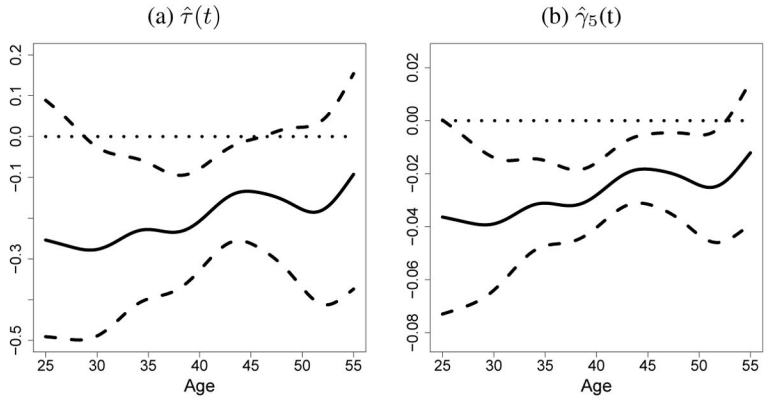

Figure 4.

(a) Estimated time-varying partial association and (b) estimated coefficient function for residuals. For each panel, the solid curve shows the estimate, the dotted line marks zero, and the dashed curves show the 2.5 and 97.5 percentiles of 500 bootstrap samples in (a) and the estimated 95% pointwise confidence band in (b).

Table 1

Results for the first stage fit

Variable	$\hat{\beta}$	95% CI
PART	0.044	(−0.056, 0.144)
HCT	0.120	(0.008, 0.232)
MCV	0.124	(0.075, 0.173)
PLAT	0.013	(0.007, 0.019)
CESD scale	−0.126	(−0.163, −0.089)