Statistica Sinica **27** (2017), 1281-1298
doi:https://doi.org/10.5705/ss.202015.0070

GENERALIZED PARTIAL LINEAR MODEL WITH UNKNOWN LINK AND UNKNOWN BASELINE FUNCTIONS FOR LONGITUDINAL DATA

Huazhen Lin[1], Ling Zhou[1] and Binhuan Wang[2]

[1]*Southwestern University of Finance and Economics and* [2]*New York University*

Abstract: In this paper we develop a generalized partial linear model for longitudinal data. In the model, we allow the link and baseline functions to be unknown. We explicitly express the estimators of regression parameters and the baseline function; hence, the computation and programming of our estimators are simple. We show that the proposed estimators of regression parameters and the baseline function are asymptotically normal with a simple variance estimator for the baseline function. In simulation studies, we demonstrate that the proposed nonparametric method is robust with limited loss of efficiency.

Key words and phrases: Generalized partial linear models, kernel method, longitudinal data, unknown baseline function, unknown link function.

1. Introduction

Longitudinal studies are often conducted in epidemiology, social science, and other biomedical research areas. To avoid the risk of misspecifying the baseline function, Moyeed and Diggle (1994), and Zeger and Diggle (1994) proposed a semiparametric model that related a response $Y(t)$ at time t to a p-dimension vector of covariates $\mathbf{X}(t)$ via the equation

$$Y(t) = V(t) + \mathbf{X}(t)'\boldsymbol{\beta} + \varepsilon(t), \tag{1.1}$$

where $V(t)$ is an unspecified smooth baseline function of t, $\boldsymbol{\beta}$ is a vector of unknown regression coefficient, and $\varepsilon(t)$ is a zero-mean Gaussian process. The model (1.1) and its variations have received a lot of attention given their flexibility and explanatory power. Martinussen and Scheike (1999, 2001), Cheng and Wei (2000), and Lin and Ying (2001) proposed estimation procedures under the formation of point processes without a specific parametric error structure. Fan and Li (2004) proposed an estimation procedure for the baseline function using a local polynomial regression and a penalized quadratic loss procedure to select significant variables. Chen and Jin (2006) proposed a least-squares-type estimator of the slope parameter via a piecewise local polynomial approximation to the nonparametric component, and the estimator was shown to be efficient when the error followed a multivariate normal distribution.

For non-normal responses, including binary, Poisson, gamma, and inverse Gaussian responses, the traditional partial linear model is not appropriate, and thus the generalized linear model is adopted and extended (McCullagh and Nelder (1989)), where a link function is introduced. By adding a parametric canonical link, Lin and Carroll (2001) and Wang, Carroll and Lin (2005) considered a generalized partial linear model for longitudinal data through a specification of the link function and proposed a kernel generalized estimation equation to solve for estimates. Lin and Carroll (2006) further proposed a backfitting method to estimate regression coefficients and baseline function with a known link function.

However, specific link functions may be inadequate in some cases (Lord (1980); Wainer (1983)) and the misspecification of link functions leads to biased estimators of regression coefficients or baseline function. The importance of choosing a correct link function has been discussed in the literature. Aranda-Ordaz (1981) and Scallan, Gilchrist and Green (1984) showed that generalized linear models can be extended to the cases with a class of parametric link functions. Weisberg and Welsh (1994), Carroll et al. (1997), Chiou and Muller (1998), Horowitz (2001), and Horowitz and Mammen (2007) proposed approaches to estimate the link function. In other contexts, the model with an unspecified link function is known as the nonparametric single-index model (Hardle, Hall and Ichimura (1993)). However, the semiparametric methods that allow unknown link functions focus on models, where no baseline function is involved.

Since the misspecification of a link function can lead to biased estimators of regression coefficients or baseline function, a nonparametric link function becomes a necessary device to analyze effects of covariates. In this paper, we propose a method to estimate regression coefficients and the baseline function when the link function is unknown.

The body of the paper is organized as follows. In Section 2, we develop the estimators of regression coefficients and baseline function. Asymptotic properties of the estimators are derived in Section 3. In Section 4, we provide a bandwidth selection procedure. Simulation results on the robustness and efficiency of the proposed estimators are presented in Section 5. In Section 6, we present a data analysis. A concluding discussion is provided in Section 7. All conditions and proofs are deferred to Supplementary Material.

2. Model and Estimation

Let $Y_{ij} = Y_i(t_{ij})$ be an outcome random variable, and $\mathbf{X}_{ij} = \mathbf{X}_i(t_{ij})$ be a $p \times 1$ vector of fixed covariates at time t_{ij} for individual i at the jth observation, where $i = 1, \cdots, n$ and $j = 1, \cdots, n_i$. We assume that $\max_i\{n_i\} < \infty$. It is well-known that the asymptotic theory of longitudinal data analysis depends on the formulation of how the data were collected. Following Wu and his collaborators (see, e.g., Hoover et al. (1998); Wu, Chiang and Hoover (1998)), we also assume that the time points t_{ij}'s are a random sample from a certain population T. By allowing the distribution of T to be any smooth function, we are not actually putting any assumption restriction on values of t_{ij}, and this assumption is just made for convenience. Other formulations can also be accommodated, with similar results obtained. Suppose the random vector $(Y_{ij}, \mathbf{X}'_{ij}, t_{ij})$ has the same distribution as that of $(Y, \mathbf{X}', T)$. Observations are correlated when they are from the same subject, or otherwise independent. We assume that the response Y_{ij} of individual i is related to covariates $\mathbf{X}_{ij}$ and t_{ij} via the generalized partial linear model

$$E\left\{Y_{ij}|\mathbf{X}_{ij}, t_{ij}\right\} = m\left\{V(t_{ij}) + \mathbf{X}'_{ij}\boldsymbol{\beta}\right\}, \qquad (2.1)$$

where $m^{-1}(\cdot)$ is an unknown link function, $V(\cdot)$ is an unknown baseline function, and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Obviously, (2.1) reduces to the partial linear model (1.1) when $m(x) = x$; (2.1) reduces to the marginal semiparametric generalized linear model (Lin and Carroll (2001); Wang, Carroll and Lin (2005); Lin and Carroll (2006)) when $m(\cdot)$ is specified by a parametric form. In addition, (2.1) only specifies the mean function of Y_{ij}, hence it can be regarded as an extension of the nonparametric single-index model (Hardle, Hall and Ichimura (1993)) with a nonparametric baseline function.

Model (2.1) continues to hold if $m(\cdot)$ and $V(\cdot)$ are replaced by $m(\cdot + c)$ and $V(\cdot) - c$ for any constant c. It also holds if $m(\cdot)$, $V(\cdot)$ and $\boldsymbol{\beta}$ are replaced by $m(\cdot/c)$, $cV(\cdot)$ and $c\boldsymbol{\beta}$ for any nonzero constant c. Therefore, scale and location normalization are required to make the model identifiable. We assume that $V(t_0) = 0$ and the first component of $\boldsymbol{\beta}$ is fixed at a certain value, where t_0 is any given constant. Thus the actual number of parameters of $\boldsymbol{\beta}$ to be estimated is reduced by one. For ease of exposition, we abuse the notation to denote the subvector of parameters to be estimated, although its first element does not need to be estimated.

First, we propose an estimation procedure for $V(\cdot)$ when $\boldsymbol{\beta}$ is given. We let $Z_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta}$, $Z = \mathbf{X}'\boldsymbol{\beta}$, $\mu(z, t) = E(Y|Z = z, T = t)$, and $p(z, t)$ be the joint density function of (Z,T). Assume that $m(\cdot)$, $\mu(\cdot,\cdot)$ and $V(\cdot)$ are differentiable with respect to all their arguments. Define $\mu_1(z,t) = \partial\mu(z,t)/\partial z$, $\mu_2(z,t) = \partial\mu(z,t)/\partial t$ and $v(t) = dV(t)/dt$. Since $\mu(z,t) = m\{V(t)+z\}$, we have

$$\mu_1(z,t) = \dot{m}\{V(t)+z\} \quad \text{and} \quad \mu_2(z,t) = \dot{m}\{V(t)+z\}\,v(t),$$

where $\dot{m}(v) = dm(v)/dv$; hence, $\mu_2(z,t)p(z,t) = v(t)\mu_1(z,t)p(z,t)$. Replacing z with Z_{ij} and making a summation over all observations, we obtain that

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_2(Z_{ij},t)p(Z_{ij},t) = v(t)\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_1(Z_{ij},t)p(Z_{ij},t).$$

Therefore,

$$v(t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_2(Z_{ij},t)p(Z_{ij},t)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_1(Z_{ij},t)p(Z_{ij},t)}. \tag{2.2}$$

Integrating both sides of (2.2) gives

$$V(t) = \int_{t_0}^{t}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_2(Z_{ij},u)p(Z_{ij},u)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mu_1(Z_{ij},u)p(Z_{ij},u)}du. \tag{2.3}$$

Expression (2.3) forms the basis of the proposed estimator of $V(\cdot)$. Throughout the paper, $0/0 = 0$.

From (2.3), it is clear that estimates of $p(z,t)$ and derivatives of $\mu(z,t)$ are necessary to derive an estimator of $V(\cdot)$ when the value of $\boldsymbol{\beta}$ is given. We estimate $\mu(z,t)$ by

$$\mu_n(z,t) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}K_1\left((Z_{ij}-z/h_1)\right)K_2\left((t_{ij}-t)/h_2\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n_i}K_1\left((Z_{ij}-z)/h_1\right)K_2\left((t_{ij}-t)/h_2\right)}, \tag{2.4}$$

where K_1 and K_2 are bounded and symmetric functions with the support $[-1,1]$, orders of r_1 and r_2, and bandwidths h_1 and h_2, respectively. Since $\mu_1(z,t) = \partial\mu(z,t)/\partial z$ and $\mu_2(z,t) = \partial\mu(z,t)/\partial t$, we obtain estimators of $\mu_1(z,t)$ and $\mu_2(z,t)$ by differentiating $\mu_n(z,t)$ with respect to z and t, respectively:

$$\begin{aligned}\mu_{1n}(z,t) &= \frac{\partial\mu_n(z,t)}{\partial z},\\ \mu_{2n}(z,t) &= \frac{\partial\mu_n(z,t)}{\partial t}.\end{aligned} \tag{2.5}$$

We estimate $p(z,t)$ by a kernel estimator:

$$p_n(z,t) = \frac{1}{Nh_1h_2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}K_1\left(\frac{Z_{ij}-z}{h_1}\right)K_2\left(\frac{t_{ij}-t}{h_2}\right), \tag{2.6}$$

where $N = \sum_{i=1}^{n} n_i$. Finally, by (2.3), we estimate $V(t)$ by

$$V_n(t) = \int_0^t \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} \mu_{2n}(\mathbf{X}'_{ij}\boldsymbol{\beta}, u) p_n(\mathbf{X}'_{ij}\boldsymbol{\beta}, u)}{\sum_{i=1}^n \sum_{j=1}^{n_i} \mu_{1n}(\mathbf{X}'_{ij}\boldsymbol{\beta}, u) p_n(\mathbf{X}'_{ij}\boldsymbol{\beta}, u)} du. \tag{2.7}$$

Given $V(\cdot)$, (2.1) reduces to a common single-index nonparametric regression problem. We estimate $\boldsymbol{\beta}$ based on the well-known Nadaraya-Watson kernel method. Here $E(Y_{ij}|V(t_{ij}) + \mathbf{X}'_{ij}\boldsymbol{\beta} = w)$ can be estimated by

$$E_n(w) = \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} Y_{ij} K\left((V(t_{ij}) + \mathbf{X}'_{ij}\boldsymbol{\beta} - w)/h\right)}{\sum_{i=1}^n \sum_{j=1}^{n_i} K\left((V(t_{ij}) + \mathbf{X}'_{ij}\boldsymbol{\beta} - w)/h\right)}, \tag{2.8}$$

where K is a bounded and symmetric kernel function with the support $[-1, 1]$, and h is a bandwidth. Then the least squares estimate of the regression coefficient $\boldsymbol{\beta}$ is obtained as the solution to

$$s_n(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} \left\{ Y_{ij} - E_n(V(t_{ij}) + \mathbf{X}'_{ij}\boldsymbol{\beta}) \right\} = 0. \tag{2.9}$$

Substituting (2.8) into (2.9), we get

$$s_n(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{X}_{ij} \left\{ Y_{ij} - \frac{\sum_{r=1}^n \sum_{k=1}^{n_r} Y_{rk} K\left((V(t_{rk}) + \mathbf{X}'_{rk}\boldsymbol{\beta} - V(t_{ij}) - \mathbf{X}'_{ij}\boldsymbol{\beta})/h\right)}{\sum_{r=1}^n \sum_{k=1}^{n_r} K\left((V(t_{rk}) + \mathbf{X}'_{rk}\boldsymbol{\beta} - V(t_{ij}) - \mathbf{X}'_{ij}\boldsymbol{\beta})/h\right)} \right\} = 0. \tag{2.10}$$

As $V_n(t)$ and $s_n(\boldsymbol{\beta})$ do not involve the unknown link function, our method can be regarded as a direct method to estimate the coefficient and baseline function. The algorithm for estimating $\boldsymbol{\beta}$ and $V(\cdot)$ can be summarized as follows:

Step 1. Obtain an initial value of $\boldsymbol{\beta}$. Because $\mu_n(z, t)$ is an estimator of $\mu(z, t) = E[Y_{ij}|Z_{ij} = z, t_{ij} = t]$, by using least squares we obtain an initial value of $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \left\{ Y_{ij} - \frac{\sum_{r=1}^n \sum_{k=1}^{n_r} Y_{rk} K_1\left((\mathbf{X}'_{rk}\boldsymbol{\beta} - \mathbf{X}'_{ij}\boldsymbol{\beta})/h_1\right) K_2\left((t_{rk} - t_{ij})/h_2\right)}{\sum_{r=1}^n \sum_{k=1}^{n_r} K_1\left((\mathbf{X}'_{rk}\boldsymbol{\beta} - \mathbf{X}'_{ij}\boldsymbol{\beta})/h_1\right) K_2\left((t_{rk} - t_{ij})/h_2\right)} \right\} \times \mathbf{X}_{ij} = 0. \tag{2.11}$$

We denote the initial estimator of $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}$. It can be shown that $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$–consistent and asymptotically normal estimator of $\boldsymbol{\beta}$.

Step 2. Estimate $V(\cdot)$ by (2.7) with $\boldsymbol{\beta}$ replaced by its estimator.

Step 3. Update $\boldsymbol{\beta}$ by (2.10) with $V(\cdot)$ replaced by its estimator from Step 2.

Step 4. Repeat Steps 2 and 3 until successive values of $\boldsymbol{\beta}$ and $V(\cdot)$ do not differ significantly.

Step 5. The discussion on large sample properties in Section 3 and the proofs in Supplementary Material show that the optimal bandwidth h_2 and the optimal kernel K_2 to estimate $\boldsymbol{\beta}$ are different from those to estimate $V(\cdot)$. Hence, we need one extra step to update the estimate of $V(\cdot)$. Step 5 fix $\boldsymbol{\beta}$ at its estimated value from Step 4, estimate $V(\cdot)$ by (2.7), while taking the bandwidth h_2 and the kernel function K_2 to be optimal for the estimation of $V(\cdot)$. Denote the bandwidth and the kernel function at the final step to be b and $\mathcal{K}$, respectively.

The estimation procedure involves choosing kernel functions K_1, K_2, K, and $\mathcal{K}$, and bandwidths h_1, h_2, h, and b. In Steps 2-4 of the algorithm, the aim is to estimate the parameter $\boldsymbol{\beta}$; hence, the kernels and the bandwidths should be optimal for this task. In Step 5, the objective is to estimate the baseline function, and thus bandwidth and kernel, in particular b and $\mathcal{K}$, should be optimal in this respect. In this paper, we take K_1 to be a sixth-order kernel, K_2 and K to be fourth-order kernels in Steps 2-4, and $\mathcal{K}$ to be a second-order kernel for Step 5, so that the assumptions on bandwidths and kernels can be satisfied. The second-, fourth- and sixth-order kernel functions can be taken from Muller (1984). The details on the selections of bandwidths are provided in Section 4.

For each iteration, the proposed method has a closed-form expression for the estimator of $V(\cdot)$ and a simple Newton-Raphson algorithm for the estimator of $\boldsymbol{\beta}$; hence, the computation and programming are straightforward. Given $\boldsymbol{\beta}$ and $V(\cdot)$, (2.1) reduces to a common nonparametric regression problem. Thus we can use any familiar nonparametric regression method, such as the local linear regression technique (Fan and Gijbels (1996)) to estimate the link function $m(\cdot)$.

3. Large Sample Properties

Let $\widehat{\boldsymbol{\beta}}$ and $\hat{V}(t)$ be the estimators of $\boldsymbol{\beta}$ and $V(t)$. We establish asymptotic normalities for $\widehat{\boldsymbol{\beta}}$ and $\hat{V}(t)$, which are summarized in Theorems 1 and 2 under some conditions given in Supplementary Material S.1. The proofs of theorems can be found in Supplementary Material S.2-S.4. The key to the proofs is to establish an asymptotic expansion of $(\hat{\mu}_2(\mathbf{x}'\hat{\boldsymbol{\beta}}, t)\hat{p}(\mathbf{x}'\hat{\boldsymbol{\beta}}, t))/(\hat{\mu}_1(\mathbf{x}'\hat{\boldsymbol{\beta}}, t)\hat{p}(\mathbf{x}'\hat{\boldsymbol{\beta}}, t))$, the estimator of $(\mu_2(\mathbf{x}'\boldsymbol{\beta}, t)p(\mathbf{x}'\boldsymbol{\beta}, t))/(\mu_1(\mathbf{x}'\boldsymbol{\beta}, t)p(\mathbf{x}'\boldsymbol{\beta}, t))$, which is obtained by some nonparametric technique, for example, in Fan and Gijbels (1996), Horowitz (1996), or Zhou, Lin and Johnson (2009).

Let $g^{(k_1,k_2,\cdots)}(a_1, a_2, \cdots) = (d^{(k_1+k_2+\cdots)}g(a_1, a_2, \cdots))/(da_1^{k_1}da_2^{k_2}\cdots)$, $Z =$

$\mathbf{X}'\boldsymbol{\beta}$ and $W = V(T) + \mathbf{X}'\boldsymbol{\beta}$. Let $p(z)$ and $p(z,t)$ be the density functions of Z and (Z,T), respectively, and $f(\cdot|w)$ be the conditional density function of T given $W = w$. Take

$$\begin{aligned}
q(w) &= E(\mathbf{X}|W = w), \;\; q(w,t) = E(\mathbf{X}|W = w, T = t),\\
\varrho(z,t) &= E[\mathbf{X}|Z = z, T = t],\\
\kappa(t) &= E\left\{m^{(1)}(W)\left[q(W) - q(W,t)\right]f(t|W)\right\}, \quad \rho(t) = E\mu_1(Z,t)p(Z,t),\\
\xi(z,t) &= -\left\{\frac{\rho^{(1)}(t)p(z)}{\rho(t)} + v(t)p^{(1)}(z) - \left[\frac{p^{(01)}(z,t) - v(t)p^{(10)}(z,t)}{p(z,t)}\right]p(z)\right\}\\
&\quad \times E\left\{m^{(1)}(W)\left[q(W) - \mathbf{X}\right]I(T \geq t)\right\},\\
\pi(t) &= \int_0^t \frac{1}{\rho(s)}E\left[p(Z,s)\left\{\varrho^{(10)}(Z,s)\mu_2(Z,s) - \varrho^{(01)}(Z,s)\mu_1(Z,s)\right\}\right]ds,\\
\mathbf{B} &= E\left\{m^{(1)}(W)\left[q(W) - \mathbf{X}\right]\pi'(T) - m^{(1)}(W)\mathbf{X}\left[q(W) - \mathbf{X}\right]'\right\} \quad \text{and}\\
\boldsymbol{\Sigma} &= \mathbf{B}^{-1}E\left\{\{Y - m(W)\}\left[\mathbf{X} - q(W) - \frac{\xi(Z,T)}{\rho(T)} - \frac{\kappa(T)p(Z)}{\rho(T)}\right]\right\}^{\otimes 2}(\mathbf{B}')^{-1}.
\end{aligned}$$

Theorem 1. *Under Conditions* 1-5 *in Supplementary Material S.*1*, if* $nh_1^{2r_1} \to 0$, $nh_2^{2r_2} \to 0$ *and* $nh^{2r_0} \to 0$ *as* $n \to \infty$*, where* r_0 *is the order of kernel* $K(\cdot)$*, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}). \tag{3.1}$$

Basically, our conditions require that to estimate $\boldsymbol{\beta}$ at the rate $n^{-1/2}$, one must undersmooth the nonparametric part. The necessity of undersmoothing to obtain an usual rate of convergence is standard in kernel literature and has analogs in spline literature (Hastie and Tibshirani (1990); Carroll et al. (1997)). On the other hand, the leading terms of $\boldsymbol{\Sigma}$ in Theorem 1 do not depend on bandwidths, indicating that these bandwidths are not crucial for the asymptotic performance of proposed estimators. The details on the selections of bandwidths and kernels are given in the next section.

Now we derive the asymptotic normality for $\hat{V}(t)$, for which extra notation is required. Let

$$\begin{aligned}
\varphi_1(t_1,t_2) &= E\left[\mu^{(01)}(Z,t_2)\xi(Z,t_2)|T = t_1\right]p_2(t_1),\\
\varphi_2(t) &= E\left[\mu^{(02)}(Z,t)\xi(Z,t)|T = t\right]p_2(t),\\
\varsigma(t) &= \frac{E\left[\mu^{(01)}(Z,t)p^{(01)}(Z,t) + \mu^{(02)}(Z,t)p(Z,t)/2\right]}{\rho(t)}
\end{aligned}$$

$$+\int_0^t \frac{\varphi_1^{(10)}(s,s)+\varphi_2(s)/2}{\rho(s)}ds,$$

$$\eta(t) = \frac{E\left[H(Z,t)p(Z,t)p(Z)\right]}{\rho^2(t)},\ H(z,t) = \mathrm{Var}(Y|Z=z,T=t),$$

$a_0 = \int x^2\mathcal{K}(x)dx,\ b_0 = \int_x \mathcal{K}^2\left(x\right)dx$, and $p_2(t)$ be the density function of T.

Theorem 2. *Under Conditions* 1-5 *in Supplementary Material* S.1*, with* $\mathcal{K}$ *a second-order kernel function, if* $\log n/\sqrt{nh_1^5} \to 0$*,* $\log n/\sqrt{nh_1^3 b^2} \to 0$*,* $nh_1^{2r_1} \to 0$ *as* $n \to \infty$*, then*

$$\sqrt{nb}\left(\hat{V}(t) - V(t) - a_0\varsigma(t)b^2\right) \xrightarrow{d} \mathcal{N}\left(0, \eta(t)b_0\right). \tag{3.2}$$

Here, the bias $\hat{V}(t) - V(t)$ is $a_0\varsigma(t)b^2$, and the variance of $\hat{V}(t) - V(t)$ is $(1/N)b\eta(t)b_0$, which can be estimated by

$$\hat{\Psi}(t) = \frac{1}{Nb}\sum_{r=1}^{n}\sum_{k=1}^{n_r}\left\{\left[\frac{\left[Y_{rk} - \hat{\mu}(\hat{Z}_{rk},t)\right]}{\hat{\rho}(t)}K_2\left(\frac{t_{rk}-t}{b}\right)\right]\hat{p}(\hat{Z}_{rk})\right\}^2, \tag{3.3}$$

where $\hat{Z}_{rk} = \mathbf{X}'_{rk}\hat{\boldsymbol{\beta}}$, $\hat{\rho}(t) = 1/N\sum_{i=1}^{n}\sum_{j=1}^{n_i}\hat{\mu}_1(\hat{Z}_{ij},t)p(\hat{Z}_{ij},t)$, $\hat{p}(z) = 1/Nh_1 \sum_{i=1}^{n}\sum_{j=1}^{n_i}K_1((\hat{Z}_{ij}-z)/h_1)$, and $\hat{\mu}(z,t)$, $\hat{\mu}_1(z,t)$ and $\hat{p}(z,t)$ are $\mu_n(z,t)$, $\mu_{1n}(z,t)$ and $p_n(z,t)$, respectively, with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}$.

Theorem 2 and its proof show that even though $\boldsymbol{\beta}$ is estimated, the asymptotic distribution of $\hat{V}$ is the same as if $\boldsymbol{\beta}$ was known. This is due to the fact that the rate of convergence of $\hat{\boldsymbol{\beta}}$ is much faster than that of $\hat{V}$: $\hat{\boldsymbol{\beta}}$ is estimated with order $O_p(n^{-1/2})$, whereas the rate of convergence of $\hat{V}$ at least is $O_p((nb)^{-1/2})$. As a consequence, the uncertainty of $\hat{\boldsymbol{\beta}}$ can be ignored. Here, $\mathcal{K}$ is a second-order kernel and $b \propto n^{-1/5}$. These selections guarantee that $V(\cdot)$ can be estimated at the optimal convergent rate, as in Zeger and Diggle (1994), Moyeed and Diggle (1994), and Fan and Li (2004) for the partial linear model with a known link function.

4. The Selections of Bandwidths

Since the leading terms in Theorems 1 and 2 do not depend on the optimal bandwidths h_1, h_2, and h, the proposed estimates are not sensitive to these bandwidths, which makes practical implementation of the proposed method much easier. From our simulations, we find that the choices of $h_1 = \sigma(\mathbf{X}'\boldsymbol{\beta})n^{-1/9}$, $h_2 = \sigma(T)n^{-1/5}$, and $h = \sigma(W)n^{-1/5}$ provide a reasonable approximation of h_1, h_2, and h, respectively, where $\sigma(Z)$ is the standard error of random variable Z.

The selection of b is crucial for the asymptotic performance of the estimator for $V(\cdot)$. We use a $K-$fold cross-validation procedure to select b that is commonly used in the literature (Tian, Zucker and Wei (2005); Fan, Lin and Zhou (2006)). Tian, Zucker and Wei (2005) and Fan, Lin and Zhou (2006) have shown empirically that the choice of the smoothing parameter can be quite flexible. Our simulations and examples also show that the cross-validation approach works well.

5. Simulations

We conducted simulation studies to assess the finite-sample performance of the proposed method. Since the validity of our method does not rely on the parametric specification of the link function, we expect our estimators of regression parameters and the baseline function to be more robust than those derived under pre-assumed parametric link functions. To investigate these issues, we conducted simulation studies and compared our method with the semiparametric linear regression model (1.1), as well as the generalized partial linear model (2.1) with a binary outcome. We assessed the performance of various estimators in terms of bias, standard error (SE), and the squared root of mean square error (RMSE).

According to our limited experience in simulations, we found that searching suitable bandwidths requires extensive computation in Steps 2 and 3. We advise using the empirical formulae $h_1 = \sigma(\mathbf{X}'\boldsymbol{\beta})n^{-1/9}$, $h_2 = \sigma(T)n^{-1/5}$, and $h = \sigma(W)n^{-1/5}$. Those seem to work well in our simulations.

5.1. Continuous outcome

Each dataset comprised $n = 400$ subjects and $n_i = m = 5$ observations per subject over time. The covariate vector was $\mathbf{X}_{ij} = (X_{1i}, X_{2i}, X_{3ij})'$, where X_{1i} and X_{2i} were subject level covariates, and X_{1i} took value 1 for one half of the subjects and 0 for the other half, mimicking a binary treatment indicator, X_{2i} was $\mathcal{N}(2, 7)$, and X_{3ij} was a time-varying covariate. We generated X_{3ij} according to the model $X_{3ij} = 4t_{ij} + b_i$, where b_i was $\mathcal{N}(0, 2)$ and t_{ij} was uniform on $(0, 1)$. We generated Y_{ij} as

$$\begin{aligned} Y_{ij} = &\Big\{V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij}\Big\} I\Big(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij} \le c\Big) \\ &+ \frac{(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij})^3}{c^2} I\Big(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij} > c\Big), \end{aligned} \tag{5.1}$$

where $V(t) = 6\sin(\pi t)$, $\beta = (3, 2, 2.5)'$, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i,n_i})'$ was $\mathcal{N}(\mathbf{0}, \Lambda)$,

Table 1. Continuous outcome: the bias, empirical standard error (SE), and root of mean square error (RMSE) of the coefficient estimators $\hat{\boldsymbol{\beta}}$ based on the 200 simulations.

			$\hat{\beta}_2$			$\hat{\beta}_3$		
c	NP	Method	bias	SE	RMSE	bias	SE	RMSE
c=17	20%	Proposed	0.0252	0.1330	0.1354	0.0251	0.1763	0.1781
		Fan&Li	1.4157	0.1441	1.4230	1.7617	0.2470	1.7789
c=20.5	10%	Proposed	0.0599	0.1234	0.1372	0.0640	0.1729	0.1843
		Fan&Li	0.5494	0.0839	0.5558	0.6892	0.1462	0.6987
c=23	5%	Proposed	0.0685	0.1058	0.1260	0.0755	0.1486	0.1667
		Fan&Li	0.2564	0.0572	0.2627	0.3179	0.1014	0.3337

The first component of $\hat{\beta}$ is fixed for both methods.

$\Lambda = (\Lambda_{rs})$, $\Lambda_{rs} = 4\rho^{|r-s|}$ and $\rho = 0.5$. We took $c = 17, 20.5, 23$ so that $Pr\left(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij} \le c\right) \approx 0.80, 0.91, 0.95$, respectively; hence, the true model approaches the semiparametric linear model (1.1) that has an identity link function, as c increases.

We adopted the method proposed by Fan and Li (2004) as an approach to estimate the partial linear model (2.1) with an identity link function. For fair comparisons with the proposed method, we fixed the first element of Fan and Li's estimator for β at 3, the true value of β_1. Table 1 presents the bias, empirical SE, and RMSE of the proposed coefficient estimator $\hat{\boldsymbol{\beta}}$ based on 200 simulations. Working bandwidths were $h_1 = 2.7, 2.7, 2.7$, $h_2 = 0.27, 0.3, 0.3$, and $h = 0.75, 0.8, 0.8$ for $c = 17, 20.5, 23$, respectively. Table 1 also shows results from Fan and Li's method with corresponding optimal bandwidth $h = 0.39, 0.26, 0.25$ for $c = 17, 20.5, 23$, respectively, selected by minimizing empirical RMSE from several pre-specified bandwidths. From Table 1, we can see that Fan and Li's estimator is severely biased even for the case where the nonlinear probability $(Pr\left(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \varepsilon_{ij} > c\right)$, termed "NP" in Table 1, is 5%. In contrast, the proposed method is almost unbiased in all cases. Although there is less variation in Fan and Li's estimator when the nonlinear probability is 5% or 10% , the large bias compromises the RMSE. As a result, the proposed estimator has much smaller RMSE and is much better in all cases. Furthermore, Table 1 shows that the proposed method is not sensitive to the nonlinear probability of the link function.

For each simulated dataset, we also obtained estimates of the baseline function $V(\cdot)$ using the proposed approach with bandwidths $h_1 = 3.7, b = 0.065$, and Fan and Li's method with its optimal bandwidth $h = 0.12, 0.1, 0.1$ for $c = 17, 20.5, 23$, respectively. Figure 1 displays the averaged estimated base-

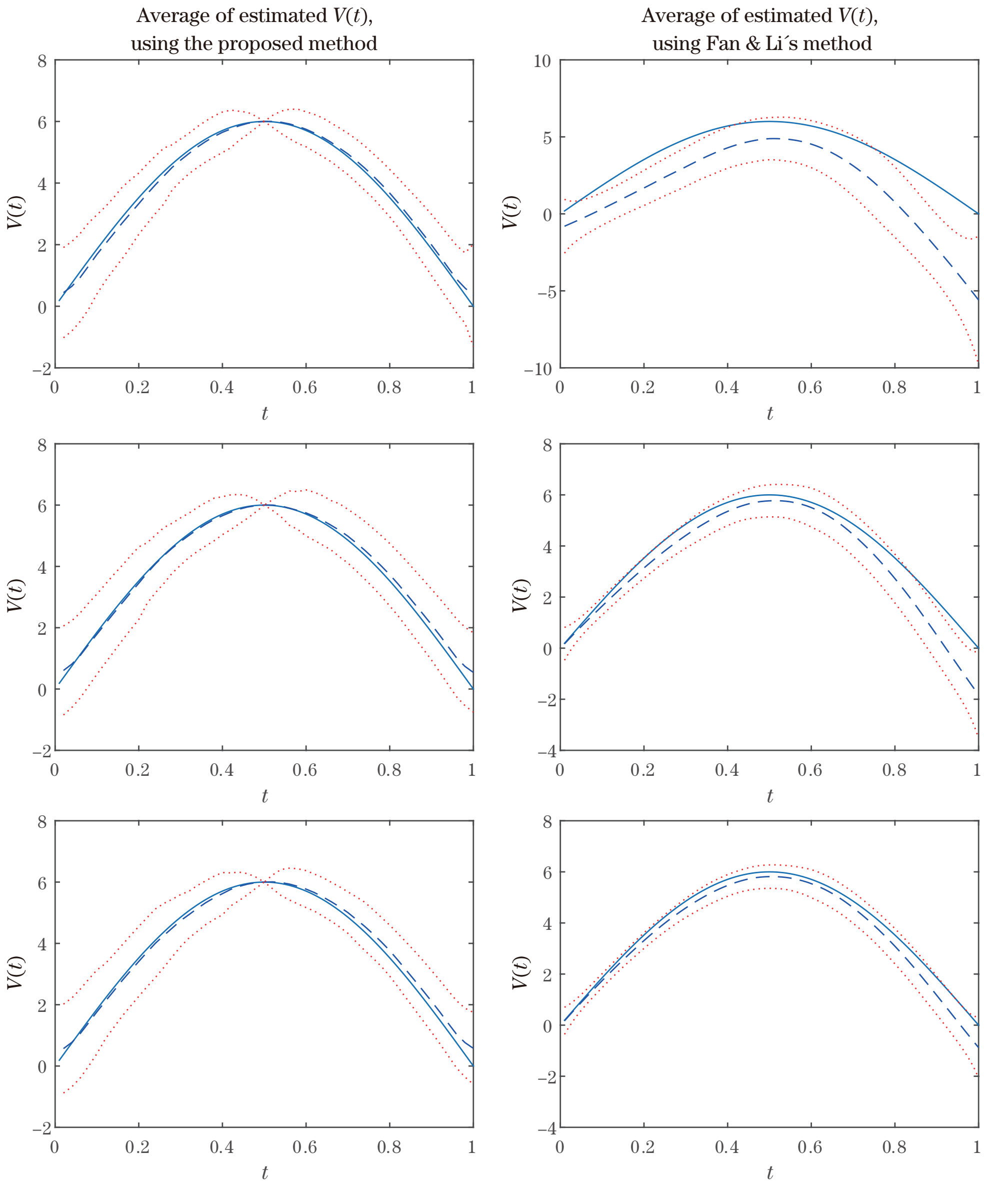

Figure 1. The first column: the true baseline function curve (solid), the average of the estimated baseline function curve (dashed) and its 95% confidence limit (dotted) over 200 replications, using the proposed method, where the estimated curve is scaled so that $\hat{V}(0.5) = 6$; The second column: the true baseline function curve (solid), the average of the estimated baseline function curve (dashed) and its 95% confidence limit (dotted) over 200 replications, using Fan & Li's method. Both are under an identical link with different correctness $p = NP = 0.8, 0.91, 0.95$ from top to bottom.

Table 2. Binary outcome: the bias, empirical standard error, and root of mean square error (RMSE) of the coefficient estimators $\hat{\boldsymbol{\beta}}$ based on the 200 simulations.

	$\hat{\beta}_2$			$\hat{\beta}_3$		
Method	bias	SE	RMSE	bias	SE	RMSE
Proposed	0.0184	0.2561	0.2568	0.0274	0.3584	0.3595
GPLM-LOGI	0.6612	0.0507	0.6632	0.8329	0.0810	0.8368

The first component of $\hat{\beta}$ is fixed at 3 for both methods.

line function and their 95% empirical pointwise confidence limits, based on 200 simulated datasets. The first column of Figure 1 shows that the proposed estimate of the baseline function is very close to the true baseline function. In contrast, the second column of Figure 1 shows that the estimate of the baseline function based on Fan and Li's method is biased, and consequently, its 95% empirical point-wise confidence interval does not cover the true curve.

5.2. Binary outcome

We generated data in the same manner as in the previous subsection, except that

$$Y_{ij} = I(V(t_{ij}) + \mathbf{X}'_{ij}\beta + \epsilon_{ij} > 12),$$

where ϵ_{ij} independently follow a mixture of two normal distributions as $0.5\mathcal{N}(2.5, 1) + 0.5\mathcal{N}(-2.5, 1)$. Roughly, $Pr\,(Y_{ij} = 1) \approx 0.42$. Therefore, the true model is the generalized partial linear model (2.1) with a binary outcome that has a non-logistic link function.

We compared our method with a generalized partial linear model with the logistic link (GPLM-LOGI), that is the most commonly used model for binary response. The R package "gplm" is used to fit the GPLM-LOGI with spline bases with 7 as the number of degrees of freedom. The number 7 was selected by minimizing the empirical RMSE from several prespecified integers. For fair comparisons with the proposed method, we also fixed the first element of β at 3, the true value. Table 2 presents the bias, empirical SE, and RMSE of the coefficient estimator $\hat{\boldsymbol{\beta}}$ using the GPLM-LOGI and the proposed method with bandwidths $h_1 = 3.3$, $h_2 = 0.27$ and $h = 0.85$, based on 200 simulations. From Table 2, we can see that the GPLM-LOGI estimator is severely biased and similar conclusions with those from Table 1 can be claimed: our method has much less bias and consequently less RMSE than the GPLM-LOGI estimator.

For each simulated dataset, we also obtaind estimates of the baseline function $V(\cdot)$ using the proposed approach with bandwidths $h_1 = 3.3, b = 0.12$, and

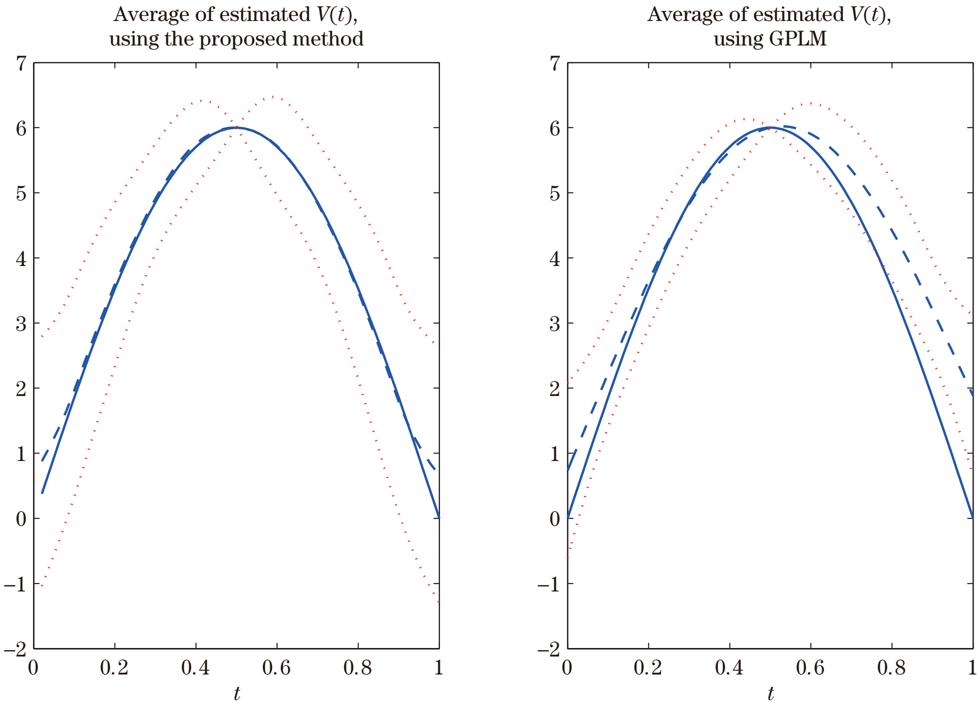

Figure 2. The first column: true baseline function curve (solid), average of estimated baseline function curve (dash) and its 95% confidence limit (dotted) over 200 replications, using the proposed method, where the estimated curve is scaled so that $\hat{V}(0.5) = 6$; The second column: true baseline function curve (solid), average of estimated baseline function curve (dash) and its 95% confidence limit (dotted) over 200 replications, using GPLM with the logistic link.

the GPLM-LOGI method. Figure 2 displays the averaged estimated baseline function and its 95% empirical pointwise confidence limits, based on 200 simulated datasets. Similar conclusions with those from Figure 1 can be obtained: our estimate of the baseline function is very close to the true baseline function, while the estimate of the baseline function based on the GPLM-LOGI method is biased, and consequently its 95% empirical point-wise confidence band does not cover the true curve.

6. Data Analysis

Multiple sclerosis (MS) is a disease that destroys the myelin that surrounds the nerves. A clinical trial on MS was conducted at the University of British Columbia, involving a drug (Betaseron) treatment with three levels (placebo, low dose, and high dose). There was a total of 50 patients in this study, wherein

17, 17 and 16 patients were randomized into placebo, low-dose, and high-dose groups, respectively. According to the trial protocol, the response variable of exacerbation was scheduled to be observed over a period of 102 weeks, with value 1 meaning the exacerbation began since the previous magnetic resonance imaging (MRI) scan, and 0 otherwise. In addition, a baseline covariate *expanded disability status scale* (EDSS) score (X_3) was collected from each patient. The central questions were whether and how the risk of exacerbation varied as a function of the dose levels and the EDSS, and how the risk of exacerbation varied with time.

This dataset was previously analyzed by Dyachkova, Petkau and White (1997) using Liang and Zeger's generalized estimating equation approach, in which the effect of the time was assumed to be constant with a specified link. However, a plot of the empirical percentage of exacerbation against time showed a very strong time-dependent relationship that could not be simply depicted by a polynomial function. Lin, Song and Zhou (2007) used a varying-coefficient logistic model to address the population-averaged relation between the probability of exacerbation and the time-varying effects of the covariates. Lin *et al.*'s results suggest the coefficients of (X_1, X_2, X_3) being constants, where $X_1 = 1$ if the treatment was a high dose and otherwise 0, and $X_2 = 1$ if the treatment was a low dose and otherwise 0. We applied the proposed generalized partial linear model with an unknown baseline function to explore the time trend and investigate whether and how the risk of exacerbation varies with the dose levels and the EDSS.

Let Y_{ij} be the indicator of exacerbation and t_{ij} be the observation time for subject i at the jth observation. The model with an unknown baseline function $V(\cdot)$ and an unknown link function $m(\cdot)$ takes the form

$$E\left\{Y_{ij} \mid \mathbf{X}_i, t_{ij}\right\} = m\left\{V(t_{ij}) + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}\right\}.$$

Take $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$. For identification, we suppose $\|\boldsymbol{\beta}\| = 1$ and the first element of $\boldsymbol{\beta}$ is fixed at a given value for the scales of $\boldsymbol{\beta}$. Although the given scales are different, the resulting estimators using the two identification conditions provide the same direction of $\boldsymbol{\beta}$, which is the aim of our procedure.

We again apply the empirical formulae $h_1 = 5 \times \sigma(X'\tilde{\beta})n^{-1/9}$, $h_2 = \sigma(T)n^{-1/5}$, and $h = 10 \times \sigma(\tilde{W})n^{-1/5}$, where $\tilde{\beta}$ and $\tilde{W}$ were calculated based on the GPLM with the logistic link function. The bandwidth $b = 44$ was chosen by a 5-fold cross-validation (Hoover et al. (1998); Fan, Lin and Zhou (2006)). The calcula-

Table 3. The estimates and SE for $\boldsymbol{\beta}$.

	Proposed		
	Estimate	SE	p-value
β_1	-0.858	0.225	0.0001
β_2	-0.234	0.448	0.6014
β_3	0.458	0.251	0.0680

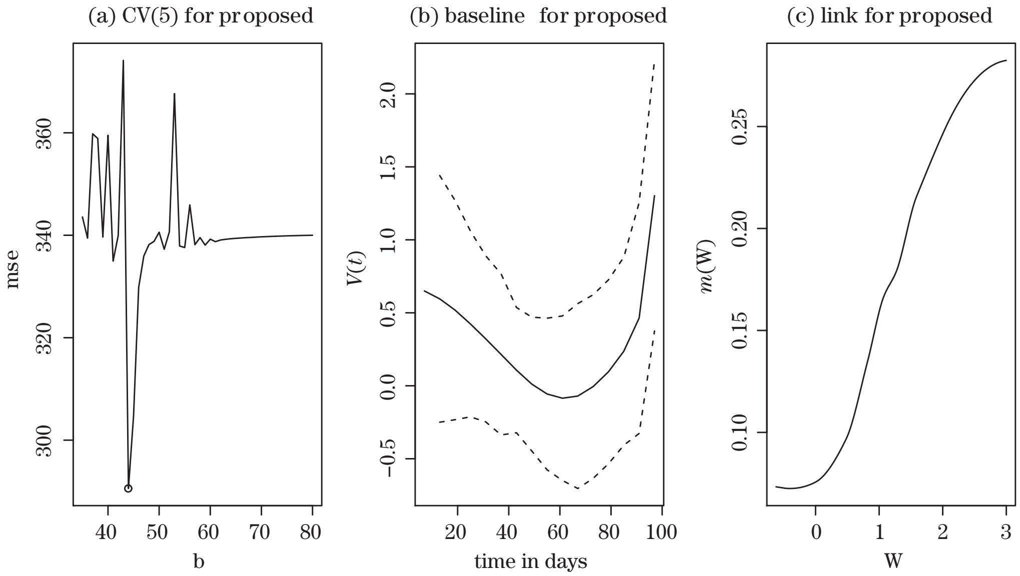

Figure 3. (a) Plot of the predicted error vs bandwidth b; (b) estimated baseline function (solid line) and its 95% confidence limit (dashed line) over 500 bootstrap replications; and (c) estimated link function.

tion of standard errors was carried out via a bootstrap resampling method with 500 bootstrap samples. Given the bandwidths, the resulting estimates of regression coefficients and the baseline function $V(\cdot)$ associated with their standard errors are provided in Table 3 and Figure 3, respectively.

From Figure 3, it is clear that there is a strong nonconstant and nonlinear time effect. The risk of exacerbation since the previous MRI scan decreases in the first 60 days, and then increases continuously to the highest value. The proposed coefficient estimates in Table 3 show that the high-dose treatment significantly reduces the risk of exacerbation, while the effects of low-dose and placebo treatment on a patient's exacerbation status are not significantly different. The EDSS score has a marginal effect and the patients with greater EDSS may have higher risk for exacerbation.

7. Discussion

In this paper, we develop a semiparametric generalized linear model with unknown link and baseline functions to analyze the effects of covariates and the serial trend for longitudinal data. The theoretical studies show that our estimators are asymptotically normal with standard convergent rates for parameters and the baseline function. Simulation studies show that our method is robust with limited loss of efficiency. We point out that the proposed estimator does not incorporate correlations among repeated measurements and that it is possible to improve the efficiency by incorporating an adaptive correlation structure among observations.

Supplementary Materials

The online supplementary material includes proofs of Theorems 1-2 and related conditions.

Acknowledgment

Lin's research is supported by the Chinese Natural Science Foundation (No. 11571282 and 11528102) and the Fundamental Research Funds for the Central Universities (JBK120509 and 14TD0046).

References

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68**, 357–363.

Chen, K. and Jin, Z. (2006). Partial linear regression models for clustered data. *Journal of the American Statistical Association* **101**, 195–204.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generally partially linear single-index models. *Journal of the American Statistical Association* **92**, 477–489.

Cheng, S. C. and Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika* **87**, 89–97.

Chiou, J. M. and Muller, H. (1998). Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association*, **93** 1376–1387.

Dyachkova, Y., Petkau, A. J. and White, R. (1997). Longitudinal analyses for magnetic resonance imaging outcomes in multiple sclerosis clinical trials. *J. Biophar. Statist.* **7**, 501–531.

Fan. J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman & Hall, London.

Fan. J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710–723.

Fan, J., Lin, H. and Zhou, Y. (2006). Local partial-likelihood estimation for lifetime data. *The Annals of Statistics* **34**, 290–325.

Hardle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics* **21**, 157–178.

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

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* **64**, 103–137.

Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* **69**, 499–513.

Horowitz, J. L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *The Annals of Statistics* **35**, 2589–2619.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association* **96**, 103–126.

Lin, H. Z., Song, Peter X.-K. and Zhou, Q. (2007). Varying-coefficient generalised linear models for longitudinal data. *Sankhya* **69**, 582–615.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045-1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society B* **68**, 69-88.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Erlbaum Associates, New Jersey.

Martinussen, T. and Scheike, T. H. (1999). A Semiparametric additive regression model for longitudinal data. *Biometrika* **86**, 691-702.

Martinussen, T. and Scheike, T. H. (2001). Sampling-adjusted analysis of dynamic additive regression models for longitudinal data. *Scandinavian Journal of Statistics* **28**, 303-323.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman & Hall, London.

Moyeed, R. A. and Diggle, P. J. (1994). Rates of convergence in semiparametric modelling of longitudinal data. *Australian Journal of Statistics* **36**, 75-93.

Muller, H. G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics* **12**, 766-774.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Scallan, A., Gilchrist, R. and Green, M. (1984). Fitting parametric link functions in generalised linear models. *Computational Statistics and Data Analysis* **2**, 37-49.

Tian, L. Zucker, D. and Wei, L. J. (2005). On the cox model with time-varying regression coefficients. *Jour. Ameri. Statist. Assoc.* **100**, 172-183.

Wainer, H. (1983). Pyramid power: searching for an error in test scoring with 830,000 helpers. *The Amercian Statistician* **37**, 87-91.

Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/Clustered data. *Journal of the American Statistical Association* **100**, 147-157.

Weisberg, S. and Welsh, A. H. (1994). Estimating the missing link function. *The Annals of Statistics* **22**, 1674-1700.

Wu, C. O., Chiang, T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a time-varying coefficient model with longitudinal data. *Journal of the American Statistical Association* **50**, 689-699.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

Zhou, X. H., Lin, H. Z. and Johnson, E. (2009). Nonparametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of the Royal Statistical Society B* **70**, 1029-1047.

Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

E-mail: linhz@swufe.edu.cn;

Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

E-mail: zhouling1003@126.com

Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA.

E-mail: Binhuan.Wang@nyumc.org

(Received February 2015; accepted June 2016)