Journal of Statistical Planning and Inference 231 (2024) 106135

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Regression analysis of longitudinal data with mixed synchronous and asynchronous longitudinal covariates

Zhuowei Sun [a], Hongyuan Cao [a,b,*], Li Chen [c], Jason P. Fine [d]

[a] School of Mathematics, Jilin University, Changchun 130012, China
[b] Department of Statistics, Florida State University, Tallahassee, FL 32306, USA
[c] Novartis Pharmaceuticals Corp., 1 Health Plaza, East Hanover, NJ, 07936, USA
[d] Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15213, USA

ARTICLE INFO

Keywords:
Asynchronous longitudinal data
Estimating equations
Last value carried forward
Omitted longitudinal covariate
Rate of convergence

ABSTRACT

In linear models, omitting a covariate that is orthogonal to covariates in the model does not result in biased coefficient estimation. This generally does not hold for longitudinal data, where additional assumptions are needed to get an unbiased coefficient estimation in addition to the orthogonality between omitted longitudinal covariates and longitudinal covariates in the model. We propose methods to mitigate the omitted variable bias under weaker assumptions. A two-step estimation procedure is proposed to infer the asynchronous longitudinal covariates when such covariates are observed. For mixed synchronous and asynchronous longitudinal covariates, we get a parametric convergence rate for the coefficient estimation of the synchronous longitudinal covariates by the two-step method. Extensive simulation studies provide numerical support for the theoretical findings. We illustrate the performance of our method on a dataset from the Alzheimer's Disease Neuroimaging Initiative study.

1. Introduction

In linear models, the Frisch–Waugh–Lovell (FWL) theorem states the equivalence of the coefficients from the full and partial regression. Specifically, using projection matrices to make the explanatory variables orthogonal to each other will lead to the same results as running the regression with all non-orthogonal explanators included (Ding, 2021; Frisch and Waugh, 1933; Lovell, 1963, 2008). In particular, if we omit a variable orthogonal to variables in the model, we get unbiased regression coefficient estimation. Recently, this idea has been used to obtain causal treatment effect estimation for longitudinal data (Bates et al., 2022). Do similar results hold in longitudinal studies with time-dependent covariates? How do we get unbiased regression coefficient estimation with omitted longitudinal covariate? Furthermore, what if the omitted longitudinal covariate is asynchronous with the longitudinal response and other longitudinal covariates in the model?

Asynchronous longitudinal data refer to the misalignment of measurement times on two longitudinal processes within an individual. Typical examples arise in analyzing electronic health records (EHR) data, where patients' health information is collected from multiple sources. EHR may include data on an individual's demographics, medication and allergies, immunization status, laboratory test results, and billing information, among others. Due to the retrospective nature of EHR, the measurement times are collected at each clinical encounter, which can be irregular and sparse across patients and asynchronous within patients. Another example comes from the Alzheimer's Disease Neuroimaging Initiative study (ADNI), where cognitive decline metrics, such as Mini-Mental State Examination (MMSE) score, are misaligned with medical imaging measurements, such as log hazard of fractional

* Corresponding author.
E-mail address: hongyuancao@gmail.com (H. Cao).

https://doi.org/10.1016/j.jspi.2023.106135
Received 22 August 2022; Received in revised form 12 October 2023; Accepted 7 December 2023
Available online 9 December 2023

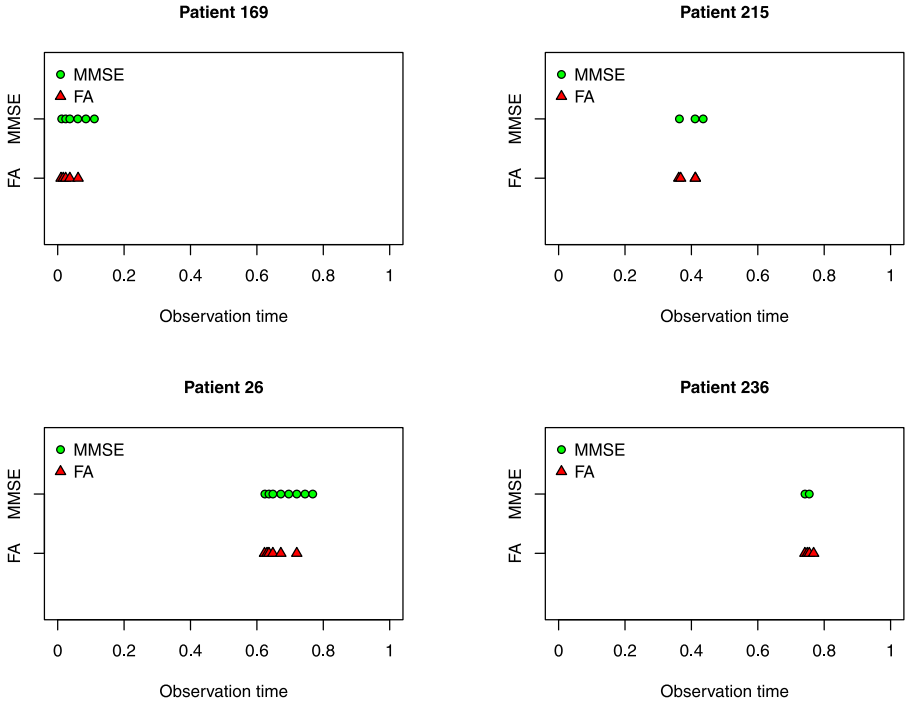

Fig. 1. Examples of individual observations.

anisotropy (FA), which reflects fiber density, axonal diameter, and myelination in white matter, within an individual. Typical measurement times of a few patients in this dataset are plotted in Fig. 1. We can see that each patient has a different number of measurements of MMSE and FA. Furthermore, the measurement times of MMSE and FA are mismatched.

For asynchronous longitudinal data, Xiong and Dubin (2010) employed a binning approach to synchronize covariates and response measurements to use existing methods for classic longitudinal data analysis. Sentürk et al. (2013) explicitly addressed the asynchronous structure for generalized varying-coefficient models with one covariate yet did not provide any theoretical justification. Cao et al. (2015) proposed a non-parametric weighting approach for generalized linear models with asynchronous longitudinal data and rigorously established inferential strategies. This was extended to a more general setup in Cao et al. (2016) and a partial linear model in Chen and Cao (2017). Recently, Li et al. (2022) studied temporally asynchronous functional imaging data, and Sun et al. (2021) examined informative measurement times for asynchronous longitudinal data. These approaches assume that all asynchronous longitudinal covariates have the same measurement times, which are asynchronous with the longitudinal response. The problem of mixed synchronous and asynchronous longitudinal covariates has not been addressed.

In this paper, we propose statistical methods for analyzing mixed synchronous and asynchronous longitudinal covariates. The longitudinal covariates have two sets, one set is measured synchronously with the longitudinal response, and another set is measured asynchronously with the longitudinal response. Suppose we are interested in inference on the synchronous longitudinal covariates and treat the asynchronous longitudinal covariates as a nuisance. Unlike classic linear models, unbiased regression coefficient estimation of the synchronous longitudinal covariates usually cannot be obtained when omitting the asynchronous longitudinal covariates, even if they are uncorrelated unless the omitted asynchronous longitudinal covariates have constant expectations over time. Ignoring this fact and only fitting synchronous longitudinal covariates with the longitudinal response may incur omitted variable bias.

To mitigate such bias, we can fit synchronous and asynchronous longitudinal covariates simultaneously like that in Cao et al. (2015). For synchronous longitudinal covariates, this one-step method implements unnecessary smoothing, which slows down the rate of convergence of the regression coefficient. To improve efficiency, we propose a two-step method. In the first step, we either fit a partial linear model of the synchronous longitudinal covariates to the longitudinal response or a linear model with centered synchronous longitudinal covariates and centered longitudinal response, omitting the asynchronous longitudinal covariates. Intuitively, we either absorb the omitted longitudinal covariates through the non-parametric intercept in the partial linear model or eliminate them through centering. We show that a parametric rate of convergence can be obtained for the regression coefficient estimation of synchronous longitudinal covariates. In the second step, residuals from the first step are fitted with the asynchronous longitudinal covariates by kernel weighting. It is established that the resulting estimator is consistent, asymptotically normal, and has the same convergence rates as that in Cao et al. (2015).

To analyze longitudinal data with partial linear models, Fan and Li (2004) developed statistical estimation and inference under the corrected specified model, whereas we are working with a misspecified model. Qian and Wang (2017) proposed a centering

approach for the analysis of classic longitudinal data assuming the model is correctly specified while we are dealing with omitted variable bias and model misspecification. Moreover, the analysis of mixed synchronous and asynchronous longitudinal covariates and omitted variable analysis for longitudinal data have not been studied before.

The rest of the paper is organized as follows. In Section 2, we elaborate on conditions for consistency of the naïve estimation of omitting the asynchronous longitudinal covariates. We then propose a partial linear model and a centering approach for consistent estimation of the regression coefficient of the synchronous longitudinal covariate and study the sampling properties of the procedure. In Section 3, we consider a two-step estimator of the regression coefficient of the asynchronous longitudinal covariate and derive its asymptotic properties and associated inferences. In addition, we derive methods and theories for analyzing synchronous and asynchronous longitudinal covariates simultaneously. In Section 4, we conduct Monte Carlo simulation studies to examine the finite sample performance of the proposed methods. Analysis of the dataset from an ADNI study illustrates the methodology in Section 5. Concluding remarks are given in Section 6. All proofs are relegated in the Supplementary Material.

2. Estimation and inference with omitted longitudinal covariates

2.1. A Naïve approach

We first look at the case where the mis-specified model is naïvely analyzed using methods from classic longitudinal data analysis omitting the asynchronous longitudinal covariates. Assume the true model is

$$Y(t) = \alpha + X(t)^T\beta + Z(t)^T\gamma + \epsilon(t), \tag{2.1}$$

where $Y(t)$ is the longitudinal outcome, α is the intercept, $X(t) \in \mathbb{R}^p$ is the observed longitudinal covariates measured synchronously with $Y(t)$, $Z(t) \in \mathbb{R}^q$ is the omitted longitudinal covariates, which may be measured asynchronously with $Y(t)$ and $X(t)$, $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ are unknown parameters to be estimated, and $\epsilon(t)$ is a mean 0 stochastic process, uncorrelated with $X(t)$ and $Z(t)$. This is a marginal model, which specifies that the conditional mean of the longitudinal response only depends on the current value of the longitudinal covariates. There is no lagged effect of the longitudinal covariates. In this subsection, our interest is on inference about the regression coefficient β. Since $Z(t)$ is omitted, in practice, we fit the misspecified model

$$Y(t) = \alpha^\diamond + X(t)^T\beta^\diamond + \epsilon^\diamond(t), \tag{2.2}$$

where $\alpha^\diamond$ is the intercept, $\beta^\diamond \in \mathbb{R}^p$ is the regression coefficient, and $\epsilon^\diamond(t)$ is a mean 0 stochastic process, uncorrelated with $\alpha^\diamond$ and $X(t)$. This naïve practice can negatively impact estimation of β in the true model (2.1).

Suppose we have a random sample of n subjects and for the ith subject, there are M_i longitudinal observations. Denote $Y_{ij} \in \mathbb{R}$ and $X_{ij} \in \mathbb{R}^p$ as the synchronous longitudinal response and covariates observed at times $t_{ij}, i = 1, \ldots, n; j = 1, \ldots, M_i$. We minimize the least square error under model (2.2)

$$\sum_{i=1}^{n}\sum_{j=1}^{M_i}(Y_{ij} - \alpha^\diamond - X_{ij}^T\beta^\diamond)^2.$$

We have

$$\hat{\beta}_n = \Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}(Y_{ij} - \alpha^\diamond). \tag{2.3}$$

Taking the expectation, we have

$$\begin{aligned}
E(\hat{\beta}_n) &= E\left\{\Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}(Y_{ij} - \alpha^\diamond)\right\} \\
&= \beta + E\left\{\Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}(\alpha + Z_{ij}^T\gamma + \epsilon_{ij} - \alpha^\diamond)\right\} \\
&= \beta + E\Bigg(\Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}\left[\alpha + E(Z_{ij})^T\gamma - \alpha^\diamond\right. \\
&\quad \left.+\{Z_{ij} - E(Z_{ij})\}^T\gamma\right]\Bigg) \\
&= \beta + E\left[\Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}\left\{\alpha + E(Z_{ij})^T\gamma - \alpha^\diamond\right\}\right] \\
&\quad + E\left[\Big(\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}X_{ij}^T\Big)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} X_{ij}\left\{Z_{ij} - E(Z_{ij})\right\}^T\gamma\right] \\
&= \beta + I + II.
\end{aligned} \tag{2.4}$$

In (2.4), I and II in general do not vanish and $\hat{\beta}$ from (2.3) is biased. We next show that the naïve approach can still work under the following conditions.

(C1) $E\{Z(t)\}$ is constant $\forall t$.

(C2) $X(t)$ and $Z(t)$ are uncorrelated $\forall t$.

Theorem 1. *Under (C1) and (C2), estimation of β in (2.1) under the mis-specified model (2.2) is unbiased.*

We remark that the intercept in (2.1) cannot be consistently estimated under the misspecified model (2.2). The proof of Theorem 1 is relegated in the Supplementary Material. Furthermore, we corroborate Theorem 1 empirically by simulation studies in Section 4.

2.2. A partial linear model approach

For general cases, instead of working with (2.2), we propose to use a partial linear model

$$Y(t) = \alpha(t) + X(t)^T\beta_p + \epsilon_p(t), \tag{2.5}$$

where $Y(t)$ is the longitudinal response, $\alpha(t)$ is the non-parametric intercept, $X(t) \in \mathbb{R}^p$ is the longitudinal covariate, $\beta_p \in \mathbb{R}^p$ is the regression coefficient and $\epsilon_p(t)$ is a mean 0 stochastic process, uncorrelated with $\alpha(t)$ and $X(t)$. As shown below, fitting the misspecified model (2.5) permits unbiased estimation of β in (2.1) under weaker conditions than those for the naïve estimator in Section 2.1.

We first define some notations. For the ith subject, we observe longitudinal response and covariates $\{Y_i(T_{ij}), X_i(T_{ij})\}, j = 1, \ldots, M_i$, where $T_{ij}, j = 1, \ldots, M_i$, are the observation times for the longitudinal measurements, where M_i is finite with probability 1. We use a counting process to represent the observation times. Specifically, $N_i(t) = \sum_{j=1}^{M_i} I(T_{ij} \le t)$ counts the number of longitudinal observations up to t (Cao et al., 2015; Lin and Ying, 2001). We use t_{ij} to denote the realized value of T_{ij}.

For the estimation of $\alpha(t)$ in (2.5), Fan and Li (2004) proposes to use local linear approximation. Specifically, for t in a neighborhood of t_0, by Taylor expansion, we have

$$\alpha(t) \approx \alpha(t_0) + \dot{\alpha}(t_0)(t - t_0) := a_0 + a_1(t - t_0),$$

where the superscript dot denotes the first-order derivative. Let $K(\cdot)$ be a kernel function and let h be a bandwidth. We aim to find $(\hat{a}_0, \hat{a}_1)$ minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{M_i} K_h(t_{ij} - t_0)\{Y_i(t_{ij}) - a_0 - a_1(t_{ij} - t_0) - X_i(t_{ij})^T\beta_p\}^2, \tag{2.6}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. The motivation for this is that we want to write $\hat{a}_0$ as a function of β_p. This is different from the GEE with working independence covariance matrix and the inverse of $K_h(t_{ij} - t_0)$ as the variance function, where the interest lies in estimating a_0, a_1 and β_p jointly (Liang and Zeger, 1986).

From (2.6), we have

$$\hat{a}_0 = \frac{\sum_{i=1}^{n}\int K_h(t - t_0)\{q_2(t - t_0) - (t - t_0)q_1(t - t_0)\}\{Y_i(t) - X_i(t)^T\beta_p\}dN_i(t)}{\sum_{i=1}^{n}\int K_h(t - t_0)\{q_2(t - t_0) - (t - t_0)q_1(t - t_0)\}dN_i(t)}, \tag{2.7}$$

where

$$q_l(t - t_0) = \sum_{i=1}^{n}\int K_h(t - t_0)(t - t_0)^l dN_i(t), \quad l = 1, 2.$$

Note that $\hat{a}_0$ is linear in $Y(t) - X(t)^T\beta_p$, so we can write the estimator of β_p in a closed form. This is accomplished by concatenating the longitudinal measurements from the first subject to the last subject into a long vector. Specifically, denote $m = \sum_{i=1}^{n} M_i$, $T^* = (T_1^*, \ldots, T_m^*) := (t_{11}, \ldots, t_{nM_n})^T$. We concatenate other functions of T^*. Denote $\epsilon^* := \epsilon(T^*) = \{\epsilon(T_1^*), \ldots, \epsilon(T_m^*)\}^T$, $\alpha^* := \alpha(T^*) = \{\alpha(T_1^*), \ldots, \alpha(T_m^*)\}^T$, $X^* = \{X_1^*, \ldots, X_m^*\} := \{X_1(t_{11}), \ldots, X_n(t_{nM_n})\}^T$ and $Y^* = \{Y_1^*, \ldots, Y_m^*\}^T := \{Y_1(t_{11}), \ldots, Y_n(t_{nM_n})\}^T$. Then $\hat{\alpha}^* = S(Y^* - X^*\beta_p)$, where $Y^* - X^*\beta_p$ is an $m \times 1$ column vector, S is an $m \times m$ symmetric matrix with the ith row and jth column entry $s_{ij} = w_{ij}(\sum_{j=1}^{m} w_{ij})^{-1}$, where $w_{ij} = K_h(T_i^* - T_j^*)\{q_{i,2} - (T_i^* - T_j^*)q_{i,1}\}$, where $q_{i,l} = \sum_{j=1}^{m} K_h(T_i^* - T_j^*)(T_i^* - T_j^*)^l, l = 1, 2$. As a result, $\hat{\alpha}^*$ is an $m \times 1$ column vector. Marginally, substituting $\hat{\alpha}^*$ into

$$Y^* = \alpha^* + X^*\beta_p + \epsilon^*,$$

we obtain

$$(I - S)Y^* = (I - S)X^*\beta_p + \epsilon^*, \tag{2.8}$$

where I is the identity matrix of dimension m. By minimizing the squared error, we have

$$\hat{\beta}_p = \{X^{*T}(I - S)^T(I - S)X^*\}^{-1}X^{*T}(I - S)^T(I - S)Y^*.$$

With $\hat{\beta}_p$ estimated, we obtain residual vector $\hat{\epsilon}_i$ for the ith subject from fitting (2.8), $i = 1, \ldots, n$. Define $\hat{C} = \text{diag}\{\hat{\epsilon}_1\hat{\epsilon}_1^T, \ldots, \hat{\epsilon}_n\hat{\epsilon}_n^T\}$, and $\hat{V} = X^{*T}(I - S)^T\hat{C}(I - S)X^*$. The variance of $\hat{\beta}_p$ can be estimated by $\widehat{\text{Var}}(\hat{\beta}_p) = D^{-1}\hat{V}D^{-1}$, where $D = X^{*T}(I - S)^T(I - S)X^*$. We next show theoretical properties of $\hat{\beta}_p$. In Fan and Li (2004), it has been shown that $\hat{\beta}_p$ is a consistent estimator of β_p in (2.5). We shall show that $\hat{\beta}_p$ is consistent for β in (2.1) and derive its asymptotic distribution. We need the following assumptions.

(C3) $N_i(t)$ is independent of $\{X_i(t), Z_i(t)\}$ for each t and $E\{dN_i(t)\} = \lambda(t)dt, i = 1, \ldots, n$. Moreover, $\epsilon(t)$ is a mean 0 process, uncorrelated with $X(t)$ and $Z(t)$.

(C4) $E\int\{\tilde{X}(t)\tilde{X}(t)^T\}\lambda(t)dt$ is positive definite and bounded, where $\tilde{X}(t) = X(t) - E\{X(t)\}$. Moreover, $E\{X^{*T}(I - S)^T(I - S)X^*\}$ is positive definite.

(C5) $K(\cdot)$ is a symmetric density function satisfying $\int zK(z)dz = 0, \int z^2$
$K(z)dz < \infty$ and $\int K(z)^2dz < \infty$.

(C6) $h \to 0$ and $nh \to \infty$.

Condition (C3) requires that the observation and error processes are independent of the longitudinal covariate processes. Condition (C4) ensures identifiability and the existence of $\hat{\beta}_p$. Conditions (C5) and (C6) specify valid kernels and bandwidths.

The following theorem, established in the Supplementary Material, states the asymptotic properties of $\hat{\beta}_p$.

Theorem 2. *Under (C2)–(C6), the asymptotic distribution of $\hat{\beta}_p$ satisfies*

$$\sqrt{n}(\hat{\beta}_p - \beta_0) \xrightarrow{d} N(0, A^{-1}\Sigma_{\gamma_0}A^{-1}), \tag{2.9}$$

where γ_0 is the true value of γ in (2.1),

$$A = E\int\{\tilde{X}(t)\tilde{X}(t)^T\}\lambda(t)dt, \quad \textit{and}$$
$$\Sigma_{\gamma_0} = E\Big\{\int\tilde{X}(t)\{\tilde{Z}(t)^T\gamma_0 + \epsilon(t)\}dN(t)\Big\}^{\otimes 2}.$$

This result is different from that in Fan and Li (2004) as it is derived under the working model (2.5) to make inferences of parameters in the true model (2.1).

We further make connections between the asymptotic variance of $\hat{\beta}_p$, $\Sigma = A^{-1}\Sigma_{\gamma_0}A^{-1}$, and its estimator, $D^{-1}\hat{V}D^{-1}$. Specifically, $(I - S)X^*$ can be viewed as a realization of $\tilde{X}$, and its inner product $D = X^{*T}(I - S)^T(I - S)X^*$ reflects A, the integral for the intensity function of the counting process $\lambda(t)dt$. Additionally $\hat{\epsilon}_i$ of $\hat{C}$ in $\hat{V}$ corresponds to $\tilde{Z}_i(t)^T\gamma_0 + \epsilon_i(t)$ in Σ_{γ_0} as both are based on the residuals with $X_i(t)$ as the covariate, $i = 1, \ldots, n$.

2.3. A centering approach

The partial linear model approach uses a non-parametric function to model the omitted longitudinal covariate $Z(t)$ and its effect. A different idea is to eliminate $Z(t)$ from the model by centering $X(t)$ and $Y(t)$. Specifically, we first take the unconditional expectation of (2.1):

$$E\{Y(t)\} = \alpha + E\{X(t)\}^T\beta + E\{Z(t)\}^T\gamma,$$

and subtract it from (2.1):

$$E\{Y(t) \mid X(t), Z(t)\} - E\{Y(t)\} = \Big[X(t) - E\{X(t)\}\Big]^T\beta + \Big[Z(t) - E\{Z(t)\}\Big]^T\gamma. \tag{2.10}$$

By (C2), taking expectation conditional on $X(t)$ only, (2.10) becomes

$$E\{\tilde{Y}(t) \mid \tilde{X}(t)\} = \tilde{X}(t)^T\beta, \tag{2.11}$$

where $\tilde{Y}(t) = Y(t) - E\{Y(t)\}$. The estimate of β can be obtained through the usual linear model analysis. Our motivation and setup differ greatly from Qian and Wang (2017), where a similar centering approach is developed to analyze classic longitudinal data.

Note that in (2.11), the unknown mean processes $E\{Y(t)\}$ and $E\{X(t)\}$ need to be estimated. This can be achieved through the Nadaraya–Watson estimator (Nadaraya, 1964; Watson, 1964). Denote $m_Y(t) = E\{Y(t)\}$ and $m_X(t) = E\{X(t)\}$. We have

$$\hat{m}_Y(t_0) = \frac{\sum_{i=1}^n\int K_h(t - t_0)Y_i(t)dN_i(t)}{\sum_{i=1}^n\int K_h(t - t_0)dN_i(t)} \quad \text{and}$$
$$\hat{m}_X(t_0) = \frac{\sum_{i=1}^n\int K_h(t - t_0)X_i(t)dN_i(t)}{\sum_{i=1}^n\int K_h(t - t_0)dN_i(t)},$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function and h is the bandwidth. Let $\hat{Y}_i(t) = Y_i(t) - \hat{m}_Y(t)$ and $\hat{X}_i(t) = X_i(t) - \hat{m}_X(t)$. The estimating equation for β is

$$U(\beta) = n^{-1}\sum_{i=1}^n\int\hat{X}_i(t)\{\hat{Y}_i(t) - \hat{X}_i(t)^T\beta\}dN_i(t). \tag{2.12}$$

Solving (2.12), we obtain

$$\hat{\beta}_c = \Big\{\sum_{i=1}^{n}\int \hat{X}_i(t)\hat{X}_i(t)^T dN_i(t)\Big\}^{-1}\sum_{i=1}^{n}\int \hat{X}_i(t)\hat{Y}_i(t)dN_i(t).$$

We shall show that $\hat{\beta}_c$ is a consistent estimator of the true β in (2.1) and establish its limiting distribution. Note that

$$\hat{\beta}_c - \beta_0 = \Big\{\sum_{i=1}^{n}\int \hat{X}_i(t)\hat{X}_i(t)^T dN_i(t)\Big\}^{-1}\sum_{i=1}^{n}\int \hat{X}_i(t)\{\hat{Y}_i(t) - \hat{X}_i^T\beta_0\}dN_i(t).$$

Therefore, the variance of $\hat{\beta}_c$ can be estimated with the sandwich formula:

$$\begin{aligned}\widehat{\mathrm{Var}}(\hat{\beta}_c) = &\{\sum_{i=1}^{n}\int \hat{X}_i(t)\hat{X}_i(t)^T dN_i(t)\}^{-1}\\ &\sum_{i=1}^{n}\Big[\int \hat{X}_i(t)\{\hat{Y}_i(t) - \hat{X}_i(t)^T\hat{\beta}_c\}dN_i(t)\Big]^{\otimes 2}\\ &\{\sum_{i=1}^{n}\int \hat{X}_i(t)\hat{X}_i(t)^T dN_i(t)\}^{-1}.\end{aligned}$$

We need an additional smoothness assumption specified in (C7) below.

(C7) $E\{X(t)\}$ and $E\{Y(t)\}$ are continuous functions for any t.

The following theorem, established in the Supplementary Material, states the asymptotic properties of $\hat{\beta}_c$.

Theorem 3. *Under (C2)–(C7), the asymptotic distribution of $\hat{\beta}_c$ satisfies*

$$\sqrt{n}(\hat{\beta}_c - \beta_0) \xrightarrow{d} N(0, A^{-1}\Sigma_{\gamma_0}A^{-1}), \tag{2.13}$$

where A and Σ_{γ_0} are the same as those in Theorem 2.

We note that $\hat{\beta}_p$ and $\hat{\beta}_c$ are asymptotically unbiased, obtain parametric root n convergence rate, and have the same limiting variance. This suggests that the newly proposed two estimators should perform similarly in practice. Simulation studies reported in Section 4 further substantiate these theoretical findings. It is counter-intuitive that we get an efficient estimation of β_0 with less information, as information in $Z(t)$ is not used. This is due to the key assumption (C2). When (C2) is violated, we propose a strategy in Section 2.4.

The partial linear model approach and centering approach require bandwidth for smoothing. We experimented with various values of bandwidths in the allowable range in the simulation studies, and the results are fairly robust to the choice of bandwidth. In practice, cross-validation may be used to select bandwidth. In terms of computation, centering is faster. The trade-off is that the partial linear model approach allows us to understand the omitted $Z(t)$ through the estimated non-parametric intercept term.

2.4. A more general approach

In the previous two subsections, we showed that omitting longitudinal covariates uncorrelated with longitudinal covariates in the model will not produce bias using the proposed methods. In general, (C2) is a strong assumption. In this section, we propose a more general approach with relaxed assumptions.

Suppose we have three sets of covariate processes, $X_1(t), X_2(t)$ and $Z(t)$, where $X_1(t)$ and $X_2(t)$ are synchronous with $Y(t)$ and $Z(t)$ is asynchronous with $Y(t)$. We are interested in statistical inference of the regression coefficient of $X_1(t)$, adjusting for $X_2(t)$ and $Z(t)$. Examples arise in randomized clinical trials, where $X_1(t)$ is treatment, $X_2(t)$ is observed synchronous covariate, and $Z(t)$ is asynchronous covariate, which may not be observed in the first stage. Suppose the full model is

$$Y(t) = \alpha + X_1(t)^T\beta_1 + X_2(t)^T\beta_2 + Z(t)^T\gamma + \epsilon(t), \tag{2.14}$$

where α is the intercept, β_1 is regression coefficient of interest, β_2 and γ are nuisance parameters and $\epsilon(t)$ is a mean 0 stochastic process, uncorrelated with $X_1(t), X_2(t)$ and $Z(t)$. Let $a^{\perp b}$ denote the projection of a on the orthogonal complement of the space spanned by b. Without loss of generality, we assume that $X_2(t)$ includes the constant 1. We have a weaker assumption.

($C2^*$) $\forall t, X_1(t)^{\perp X_2(t)}$ and $Z(t)^{\perp X_2(t)}$ are uncorrelated.

Condition ($C2^*$) is the unconfoundedness assumption. The relationship between different variables at time t is depicted in Fig. 2. Under this assumption, we use ideas from the FWL theorem (Frisch and Waugh, 1933; Lovell, 1963; Ding, 2021) to convert the problem to our set-up. Specifically, we first get $X_1(t)^{\perp X_2(t)}$, which can be obtained through the residual from regressing $X_1(t)$ on $X_2(t)$. We get $Z(t)^{\perp X_2(t)}$ by the same token. Denote $H_2(t) = I - X_2(t)^T\left\{X_2(t)X_2(t)^T\right\}^{-1}X_2(t)$. Multiplying $H_2(t)$ to both sides of (2.14), we get

$$H_2(t)Y(t) = H_2(t)X_1(t)^T\beta_1 + H_2(t)Z(t)^T\gamma + \epsilon(t).$$

By ($C2^*$), $H_2(t)X_1(t)^T$ and $H_2(t)Z(t)^T$ are uncorrelated, and the proposed method would work. We can apply the method proposed in Sections 2.2 and/or 2.3 to get an unbiased estimation of β_1. Numerical support for this can be found in simulation studies in Section 4.

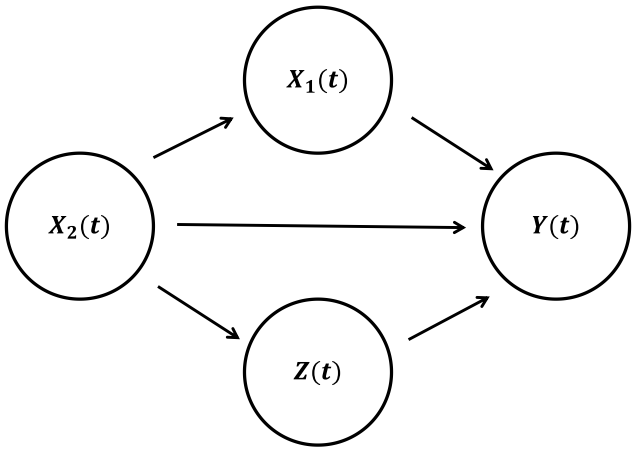

Fig. 2. Graphical representation of variables in model (2.14).

3. Estimation and inference of asynchronous longitudinal covariates

3.1. A two-step method

In this section, we consider the case that longitudinal covariates $X(t)$ and $Z(t)$ are asynchronous and longitudinal response $Y(t)$ is observed in alignment with $X(t)$. There is no existing literature to deal with such mixed synchronous and asynchronous longitudinal covariates. Specifically, suppose we have a random sample of n subjects. For subject $i = 1, \ldots, n$, $N_i(t,s) = \sum_{j=1}^{M_i} \sum_{k=1}^{L_i} I(t_{ij} \le t, s_{ik} \le s)$ counts the number of observation times up to t on $X(\cdot)$ and $Y(\cdot)$ and up to s on $Z(\cdot)$, where $t_{ij}, j = 1, \ldots, M_i$ are the observation times of $X(\cdot)$ and $Y(\cdot)$ and $s_{ik}, k = 1, \ldots, L_i$ are the observation times of $Z(\cdot)$. Denote $E\{dN_i(t,s)\} = \eta(t,s)dtds, i = 1, \ldots, n$. We propose a two-step approach to estimate β and γ in (2.1).

Step 1: Regress longitudinal response $Y(t)$ on synchronous longitudinal covariate $X(t)$ to get $\hat{\beta}$ and the residuals.

Step 2. Regress residuals from Step 1 on asynchronous longitudinal covariate $Z(t)$ to estimate $\hat{\gamma}$.

In Step 1, either a partial linear model approach or a centering approach can be used as they have the same asymptotic distribution. Once $\hat{\beta}$ is obtained, we compute the residual $\hat{\omega}_i(T_{ij}) = Y_i(T_{ij}) - X_i(T_{ij})^T\hat{\beta}$. In Step 2, to estimate γ, we propose the following estimating equation (Cao et al., 2015)

$$U^f(\gamma) = n^{-1}\sum_{i=1}^{n} \iint K_h(t-s)Z_i(s)\{Y_i(t) - Z_i(s)^T\gamma - X_i(t)^T\hat{\beta}\}dN_i(t,s), \tag{3.15}$$

where $K_h(t) = K(t/h)/h, K(t)$ is a symmetric kernel function, usually taken to be the Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ and h is the bandwidth. Solving $U^f(\gamma) = 0$, we obtain

$$\hat{\gamma} = \left\{\sum_{i=1}^{n} \iint K_h(t-s)Z_i(s)Z_i(s)^T dN_i(t,s)\right\}^{-1} \times \sum_{i=1}^{n} \iint K_h(t-s)Z_i(s)\{Y_i(t) - X_i(t)^T\hat{\beta}\}dN_i(t,s).$$

The idea of the two-step approach is intuitive. Note (2.1) can be written as $Y(t) - X(t)^T\beta = Z(t)^T\gamma + \epsilon(t)$, where we abuse notation by absorbing intercept α into β and letting the first entry of vector X_{ij} to be 1. Once we get $\hat{\beta}$, the estimation of γ can proceed as an asynchronous regression problem with residual $Y(t) - X(t)^T\hat{\beta}$ as the new response.

We next present asymptotic properties of $\hat{\gamma}$. Denote $\sigma^2(t) = \mathrm{Var}\{\epsilon(t)\}$ and let γ_0 be the true regression coefficient. We need the following conditions.

(C8) $\eta(t,s)$ is twice continuously differentiable for $(t,s) \in [0,1]^{\otimes 2}$. Moreover, For $t_1 \ne t_2, s_1 \ne s_2$, $P\{dN(t_1,s_1) = 1 \mid N(t_2,s_2) - N(t_2-,s_2-) = 1\} = f(t_1,t_2,s_1,s_2)dt_1ds_1$, where $f(t_1,t_2,s_1,s_2)$ is continuous for $t_1 \ne t_2, s_1 \ne s_2$, and $f\{t_1\pm, t_2\pm, s_1\pm, s_2\pm\}$ exists.

(C9) $E\{Z(t)Z(s)^T\}$ is twice continuously differentiable for $(t,s) \in [0,1]$. In addition,

$\int E\{Z(s)Z(s)^T\}\eta(s,s)ds$ is positive definite and

$$\| \int E\{Z(s)Z(s)^T\}\eta(s,s)\sigma^2(s)ds\|_\infty < \infty,$$

where for a square matrix A, $\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$.

(C10) $nh \to \infty$ and $nh^5 \to 0$.

The reason we impose $nh^5 \to 0$ is to eliminate bias. The bias is of order $O(h^2)$. We require the kernel function $K(\cdot)$ to be a symmetric density function satisfying $\int zK(z)dz = 0$. In the proof, the estimating equation includes the $K_h(t-s)$ term. After a change of variable, the first order term involving h is multiplied by $zK(z)$ in the integration, which vanishes.

The following theorem states the asymptotic properties of $\hat{\gamma}$.

Theorem 4. *Under (C3)–(C5), (C7)–(C10), the asymptotic distribution of $\hat{\gamma}$ satisfies*

$$\sqrt{nh}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, A_{\gamma_0}^{-1} \Sigma A_{\gamma_0}^{-1}), \tag{3.16}$$

where

$$A_{\gamma_0} = \int E\left\{Z(s)Z(s)^T\right\} \eta(s,s)ds \quad \text{and}$$
$$\Sigma = \int K(z)^2 dz \int E\left\{Z(s)Z(s)^T\right\} \eta(s,s)\sigma^2(s)ds.$$

The asymptotic distribution of $\hat{\gamma}$ is the same as that in Cao et al. (2015) with the identity link function. As $\hat{\gamma}$ has $\sqrt{nh}$ rate of convergence, slower than the $\sqrt{n}$ rate of convergence of $\hat{\beta}$, plugging in $\hat{\beta}_p$ or $\hat{\beta}_c$ does not affect the limiting distribution of $\hat{\gamma}$, which corroborates the validity of the proposed two-step method. That is, estimating γ is as efficient as if β were known *a priori*. The variance of $\hat{\gamma}$ can be estimated from the sandwich formula

$$\widehat{\mathrm{Var}}(\hat{\gamma}) = \left\{\sum_{i=1}^{n} \iint K_h(t-s)Z_i(s)Z_i(s)^T dN_i(t,s)\right\}^{-1}$$
$$\sum_{i=1}^{n}\left[\iint K_h(t-s)Z_i(s)\{Y_i(t) - Z_i(s)^T\hat{\gamma} - X_i(t)^T\hat{\beta}\}dN_i(t,s)\right]^{\otimes 2}$$
$$\left\{\sum_{i=1}^{n} \iint K_h(t-s)Z_i(s)Z_i(s)^T dN_i(t,s)\right\}^{-1}.$$

3.2. Simultaneous estimation of synchronous and asynchronous longitudinal covariates

For the mixed synchronous and asynchronous longitudinal covariates, the natural idea is to use estimating equations to get estimators of β and γ simultaneously, similar to Cao et al. (2015). Denote $W(t,s) = \{X(t)^T, Z(s)^T\}^T$. We use the estimating equation

$$U_w(\beta,\gamma)$$
$$= n^{-1}\sum_{i=1}^{n} \iint K_h(t-s)W_i(t,s)\{Y_i(t) - Z_i(s)^T\gamma - X_i(t)^T\beta\}dN_i(t,s) \tag{3.17}$$

to get $\hat{\beta}_w$ and $\hat{\gamma}_w$, respectively. In Cao et al. (2015), they studied the case that all longitudinal covariates are observed at the same times, which are asynchronous with the longitudinal response. In this section, we look at the scenario where some longitudinal covariates are synchronous with the longitudinal response, and some longitudinal covariates are asynchronous with the longitudinal response. As shown in Theorem 5, the resulting estimators are asymptotically unbiased. However, the convergence rate of the obtained estimator of β is slower than the parametric $\sqrt{n}$ rate, as obtained through the partial linear model or centering approach. The rationale is that unnecessary smoothing on $X(t)$ makes the corresponding estimator less efficient. This is further demonstrated in the simulation studies. We need the following assumption.

(C11) $\mathrm{E}\{W(t,s)W(t,s)^T\} \in \mathbb{R}^{(p+q)\times(p+q)}$ is twice continuously differentiable for $(t,s) \in [0,1]$ where $W(t,s) = \{X(t)^T, Z(s)^T\}^T$ $\int \mathrm{E}\{W(t,t)W(t,t)^T\}\eta(t,t)\,\mathrm{d}t$ is positive definite and

$$\| \int \mathrm{E}\{W(t,t)W(t,t)^T\}\sigma^2(t)\eta(t,t)\,\mathrm{d}t\|_\infty < \infty \quad \forall t,$$

where for a square matrix A, $\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$.

Theorem 5. *Under conditions (C8), (C10) and (C11), let $\hat{\theta}_w = (\hat{\beta}_w^T, \hat{\gamma}_w^T)^T$ and $\theta_0 = (\beta_0^T, \gamma_0^T)^T$, the asymptotic distributions of $\hat{\theta}_w$ satisfies*

$$\sqrt{nh}(\hat{\theta}_w - \theta_0) \xrightarrow{d} N(0, A_{\theta_0}^{-1}\Sigma_{\theta_0}A_{\theta_0}^{-1}),$$

where

$$A_{\theta_0} = \int \mathrm{E}\{W(t,t)W(t,t)^T\}\eta(t,t)\,\mathrm{d}t \quad \text{and}$$
$$\Sigma_{\theta_0} = \int K(z)^2\,\mathrm{d}z \int \mathrm{E}\{W(t,t)W(t,t)^T\}\sigma^2(t)\eta(t,t)\,\mathrm{d}t.$$

The sandwich formula can estimate the asymptotic covariance matrix

$$\widehat{\mathrm{Var}}(\hat{\theta}_w)$$

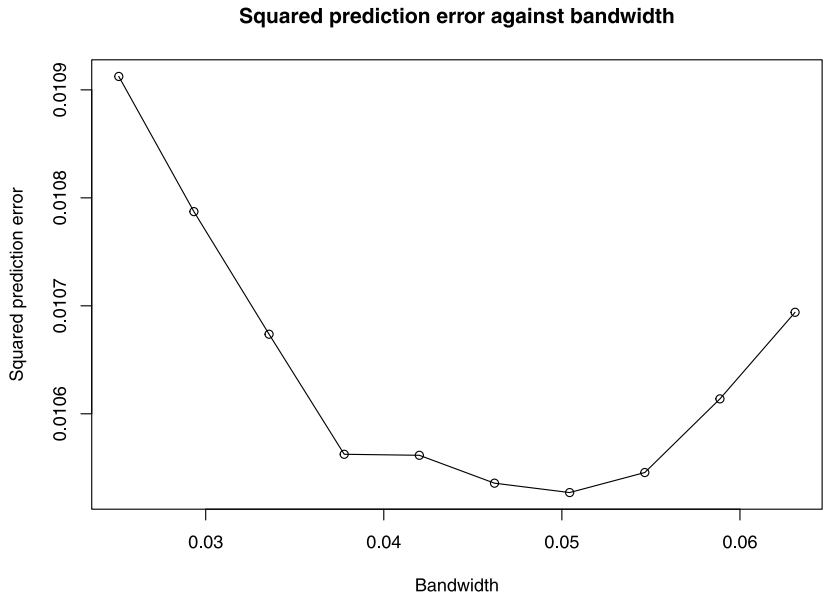

Fig. 3. Typical squared prediction error against bandwidth for $n = 100$, $E\{Z(t)\} = 2\sin(2\pi t)$, with bandwidth ranging from $n^{-0.8}$ to $n^{-0.6}$.

$$= \left\{ \sum_{i=1}^{n} \iint K_h(t-s) W_i(t,s) W_i(t,s)^T d N_i(t,s) \right\}^{-1}$$
$$\sum_{i=1}^{n} \left[\iint K_h(t-s) W_i(t,s) \{ Y_i(t) - Z_i(s)^T \hat{\gamma} - X_i(t)^T \hat{\beta} \} d N_i(t,s) \right]^{\otimes 2}$$
$$\left\{ \sum_{i=1}^{n} \iint K_h(t-s) W_i(t,s) W_i(t,s)^T d N_i(t,s) \right\}^{-1}.$$

It is worth noting that Theorem 5 requires a weaker condition on the relationship between $X(t)$ and $Z(t)$ than the two-step approach. For the latter to work, we need the key assumption specified in ($C2^*$). This reflects the trade-off between robustness and efficiency.

3.3. Bandwidth selection

Our approach to estimating γ depends on the bandwidth selection. In synchronous longitudinal data, cross-validation is usually used to select the optimal bandwidth by minimizing the squared prediction error. However, in asynchronous longitudinal data, since observations are mismatched, prediction errors are not well defined. Cao et al. (2015) proposes to minimize the mean squared error by calculating bias and variance separately. First, based on the asymptotic result, bias is in the same order as the bandwidth square. One can regress the squared bandwidth with the estimated regression coefficient to obtain the slope estimate. The bias is approximated by multiplying the slope estimate and the squared bandwidth. Second, the data are split into two halves, and coefficient estimates are obtained for each half. The squared difference of the two coefficient estimates divided by 4 approximates the variance. The mean squared error is squared bias plus variance, and the optimal bandwidth is chosen to be the one that minimizes the mean squared error. This method is somewhat *ad hoc* since only two folds of data are used, and the squared difference is coarse and may be an imprecise estimate of the variance.

We propose a new kernel-smoothed cross-validation to select the optimal bandwidth within a certain range. First, we split the data into several folds and estimated the regression coefficient without one fold. In computing the prediction error, we use kernel smoothing to deal with the mismatched response and covariate. Specifically, suppose $\hat{\beta}^{(-k)}$ and $\hat{\gamma}^{(-k)}$ are estimates without the kth fold. The squared prediction error for the kth fold is computed as

$$\frac{\sum_{i=1}^{n^{(k)}} \iint K_h(t-s) \{ Y_i(t) - X_i(t)^T \hat{\beta}^{(-k)} - Z_i(s)^T \hat{\gamma}^{(-k)} \}^2 d N_i(t,s)}{\sum_{i=1}^{n^{(k)}} \iint K_h(t-s) d N_i(t,s)},$$

where $n^{(k)}$ is the number of subjects in the kth fold. We take the average of the squared prediction errors over all folds and select the bandwidth with the smallest average squared prediction error. A typical functional relationship between the average squared prediction error and bandwidth is depicted in Fig. 3.

Table 1
1000 simulation results for inference of β with $h = n^{-0.6}$.

	Naïve				PLM				Centering			
	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
$E\{Z(t)\} = 2$												
$n = 100$	−0.002	0.207	0.189	91	−0.003	0.219	0.196	90	−0.003	0.219	0.196	90
$n = 400$	0.003	0.106	0.103	94	0.002	0.112	0.107	93	0.002	0.112	0.107	93
$n = 900$	0.002	0.073	0.070	94	0.002	0.076	0.073	93	0.002	0.076	0.073	93
$E\{Z(t)\} = 0.5 + t$												
$n = 100$	−0.053	0.207	0.188	91	0.016	0.220	0.193	90	0.016	0.220	0.193	90
$n = 400$	−0.067	0.106	0.104	89	−0.002	0.111	0.108	94	−0.002	0.111	0.108	94
$n = 900$	−0.059	0.073	0.070	85	0.005	0.076	0.073	94	0.005	0.076	0.073	94
$E\{Z(t)\} = 0.5 + t^2$												
$n = 100$	−0.060	0.217	0.190	90	0.003	0.227	0.196	90	0.003	0.227	0.196	90
$n = 400$	−0.064	0.107	0.103	89	−0.003	0.113	0.107	93	−0.003	0.113	0.107	94
$n = 900$	−0.062	0.071	0.071	86	−0.002	0.074	0.074	94	−0.002	0.074	0.074	94
$E\{Z(t)\} = 2\sin(2\pi t)$												
$n = 100$	0.237	0.226	0.204	74	−0.002	0.225	0.194	90	−0.002	0.225	0.195	90
$n = 400$	0.233	0.118	0.111	46	0.001	0.115	0.108	93	0.001	0.115	0.108	93
$n = 900$	0.231	0.079	0.076	17	0.002	0.076	0.074	94	0.002	0.076	0.074	94

Note: "Bias" is the empirical bias, "SD" is the sample standard deviation, "SE" is the average of the standard error estimates, "CP"/100 represents the coverage probability of the 95% confidence interval of $\hat{\beta}$. Naïve denotes the naive method, PLM denotes the partial linear model-based method and Centering denotes the centering method.

4. Numerical studies

In this section, we investigate the finite sample performance of the proposed estimators through Monte Carlo simulations.

4.1. Omitted longitudinal covariate

We first examine the performance of $\hat{\beta}_p$ and $\hat{\beta}_c$ along with the naïve estimator $\hat{\beta}_n$ when some important covariates are omitted. The model we use to generate data is

$$Y(t) = \alpha + X(t)^T\beta + Z(t)^T\gamma + \epsilon(t),$$

where $\alpha = 1, \beta = 2$ and $\gamma = -1$. We generate 1000 datasets, each consisting of $n = 100, 400$, or 900 subjects. Bandwidth is fixed at $n^{-0.6}$; other bandwidths yield similar results which are relegated in the Supplementary Material. The number of observations for each subject is Poisson(5)+1, and the observation times are generated from the uniform distribution $\mathcal{U}(0, 1)$. The covariate processes $X(t), Z(t)$ and the error process $\epsilon(t)$ are generated in the following manner:

1. Generate a Gaussian process $\nu(t)$ with mean 0 and $\mathrm{Cov}\{\nu(t), \nu(s)\} = e^{-|t-s|}$;
2. $Z(t) = Z'(t) + \nu(t)$, where $E\{Z'(t)\} = 0.5 + t, 0.5 + t^2, 0.5 + \sqrt{t}$ or $E\{Z'(t)\} = 2$ and $\mathrm{Cov}\{Z'(t), Z'(s)\} = e^{-|t-s|}$;
3. $X(t) = X'(t) + \omega\nu(t)$, where $\omega \sim N(0, 1)$, independent of $\nu(t)$, $E\{X'(t)\} = \sqrt{t}$, and $\mathrm{Cov}\{X'(t), X'(s)\} = e^{-|t-s|}$;
4. $\epsilon(t) = \tau\nu(t)$, where $\tau \sim N(0, 1)$, independent of $\nu(t)$.

We observe from Table 1 that when $E\{Z(t)\} = 2$, the naïve estimator $\hat{\beta}_n$ performs reasonably well. For all other time-varying $E\{Z(t)\}$, there is evidence of substantial bias and poor coverage probabilities for the true β. The partial linear model (PLM) based estimator $\hat{\beta}_p$ and centering (Centering) approach based estimator $\hat{\beta}_c$ have almost identical performance with $\sqrt{n}$ rate of convergence. As the sample size increases, the empirical and model-based standard errors tend to agree, and the coverage is close to the nominal 95% level. The performance improves with larger sample sizes.

4.2. Asynchronous longitudinal covariate

We next study coefficient estimation when $Z(t)$ is mismatched with $X(t)$ and $Y(t)$. The simulation setup is the same as that in Section 4.1 except that $X(t), Z(t)$ and the error process $\epsilon(t)$ are generated as follows. The covariate process $X(t)$ and $Z(t)$ are both Gaussian, with $E\{X(t)\} = \sqrt{t}, \mathrm{Cov}\{X(t), X(s)\} = \mathrm{Cov}\{Z(t), Z(s)\} = e^{-|t-s|}$ and $E\{Z(t)\} = 0.5 + t, 0.5 + t^2, 2\sin(2\pi t)$ or $E\{Z(t)\} = 2$. The error process $\epsilon(t)$ is Gaussian with $E\{\epsilon(t)\} = 0$ and $\mathrm{Cov}\{\epsilon(t), \epsilon(s)\} = 2^{-|t-s|}$. $X(t), Z(t)$ and $\epsilon(t)$ are independently generated. As estimators based on partial linear model approach and centering approach have the same asymptotic distribution, we illustrate the two-step method with the centering approach, denoted as Centering + KS in Table 2. For comparison, we implement three alternative estimation procedures.

Table 2
1000 simulation results for β, γ and α.

	n	LVCF				Centering+LVCF				Centering+KS				KS			
		Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
$E\{Z(t)\} = 2$																	
β	100	0.002	0.099	0.091	92	−0.005	0.130	0.116	91	−0.005	0.130	0.116	91	0.001	0.124	0.113	93
	400	0.0004	0.049	0.047	94	−0.008	0.064	0.059	93	−0.008	0.064	0.059	93	0.001	0.082	0.081	94
	900	−0.002	0.032	0.031	94	−0.008	0.040	0.040	94	−0.008	0.040	0.040	94	−0.001	0.072	0.069	93
γ	100	0.121	0.102	0.095	74	0.128	0.102	0.095	72	0.023	0.129	0.122	91	0.016	0.129	0.116	91
	400	0.118	0.051	0.049	34	0.120	0.051	0.049	33	0.006	0.084	0.085	94	0.004	0.084	0.083	95
	900	0.118	0.032	0.033	5	0.119	0.032	0.033	5	0.0002	0.071	0.072	95	−0.0005	0.071	0.072	95
α	100	−0.245	0.242	0.226	80	−0.254	0.249	0.217	76	−0.043	0.308	0.286	91	−0.032	0.305	0.273	91
	400	−0.240	0.121	0.116	47	−0.237	0.125	0.112	46	−0.006	0.193	0.194	94	−0.008	0.197	0.194	96
	900	−0.236	0.079	0.078	14	−0.234	0.081	0.075	13	0.002	0.161	0.163	94	−0.001	0.166	0.167	95
$E\{Z(t)\} = 0.5 + t$																	
β	100	−0.011	0.093	0.091	94	−0.003	0.120	0.114	92	−0.003	0.120	0.114	92	−0.002	0.122	0.113	92
	400	−0.011	0.046	0.047	95	−0.001	0.060	0.059	94	−0.001	0.060	0.059	94	−0.001	0.085	0.082	93
	900	−0.010	0.031	0.032	94	−0.001	0.040	0.039	94	−0.001	0.040	0.039	94	−0.003	0.069	0.070	95
γ	100	0.114	0.095	0.092	74	0.120	0.095	0.092	72	0.008	0.120	0.117	91	0.002	0.119	0.112	91
	400	0.119	0.047	0.047	30	0.120	0.047	0.047	28	0.004	0.082	0.081	94	0.002	0.082	0.080	95
	900	0.119	0.031	0.032	2	0.119	0.031	0.031	3	0.003	0.068	0.069	94	0.003	0.069	0.069	95
α	100	−0.230	0.150	0.148	64	−0.243	0.156	0.136	56	−0.010	0.186	0.187	94	−0.004	0.189	0.180	93
	400	−0.236	0.078	0.076	13	−0.244	0.083	0.069	9	−0.001	0.126	0.124	94	0.000	0.131	0.128	95
	900	−0.234	0.051	0.051	0	−0.241	0.054	0.047	0	0.000	0.102	0.104	94	0.002	0.107	0.109	96
$E\{Z(t)\} = 0.5 + t^2$																	
β	100	−0.021	0.094	0.091	93	−0.003	0.125	0.115	93	−0.003	0.125	0.115	93	−0.005	0.125	0.112	92
	400	−0.021	0.047	0.047	93	−0.001	0.059	0.059	95	−0.001	0.059	0.059	95	−0.001	0.083	0.081	94
	900	−0.018	0.032	0.032	91	0.003	0.040	0.040	95	0.003	0.040	0.040	95	0.003	0.071	0.069	95
γ	100	0.109	0.091	0.091	77	0.115	0.090	0.091	75	0.018	0.119	0.114	92	0.011	0.120	0.110	93
	400	0.112	0.048	0.047	33	0.113	0.048	0.047	33	0.009	0.085	0.080	92	0.008	0.085	0.079	92
	900	0.109	0.033	0.032	7	0.109	0.033	0.032	7	0.000	0.070	0.068	94	−0.001	0.070	0.068	94
α	100	−0.202	0.143	0.140	68	−0.219	0.153	0.126	56	−0.013	0.185	0.175	91	−0.006	0.189	0.169	92
	400	−0.203	0.072	0.072	20	−0.218	0.075	0.064	12	−0.008	0.119	0.114	93	−0.007	0.123	0.119	93
	900	−0.201	0.049	0.048	1	−0.216	0.051	0.043	0	−0.001	0.098	0.095	93	−0.000	0.104	0.102	94
$E\{Z(t)\} = 2\sin(2\pi t)$																	
β	100	0.087	0.111	0.111	86	0.010	0.121	0.116	94	0.010	0.121	0.116	94	0.004	0.123	0.114	93
	400	0.082	0.059	0.056	69	0.008	0.061	0.059	94	0.008	0.061	0.059	94	−0.001	0.086	0.082	93
	900	0.084	0.038	0.038	40	0.009	0.043	0.040	92	0.009	0.043	0.040	93	0.001	0.073	0.070	94
γ	100	0.257	0.059	0.056	1	0.255	0.059	0.055	1	0.009	0.064	0.061	91	0.006	0.065	0.061	93
	400	0.257	0.028	0.029	0	0.253	0.028	0.028	0	0.003	0.046	0.046	93	0.002	0.047	0.046	95
	900	0.257	0.019	0.019	0	0.253	0.019	0.019	0	0.001	0.041	0.040	93	−0.0003	0.041	0.040	94
α	100	0.279	0.154	0.143	50	0.336	0.157	0.119	25	−0.011	0.157	0.150	92	−0.006	0.158	0.147	93
	400	0.282	0.077	0.073	3	0.336	0.078	0.060	0	−0.008	0.097	0.095	93	−0.001	0.104	0.102	94
	900	0.282	0.049	0.049	0	0.338	0.050	0.040	0	−0.005	0.080	0.078	93	0.0003	0.088	0.087	94

Note: "Bias" is the empirical bias, "SD" is the sample standard deviation, "SE" is the average of the standard error estimates, "CP"/100 represents the coverage probability of the 95% confidence interval for $\hat{\beta}, \hat{\gamma}$, and $\hat{\alpha}$, respectively. LVCF denotes the last value carried forward method, Centering+LVCF denotes the two-step approach with the centering method in the first step and LVCF in the second step, Centering+KS denotes the two-step approach with the centering method in the first step and the kernel smoothing method in the second step, and KS denotes the one step kernel weighting approach.

1. Apply the last value carried forward (LVCF) to estimate β and γ simultaneously. In longitudinal studies, a naïve approach to analyzing asynchronous longitudinal data is the last value carried forward method. If data at a certain time point are missing, the observation at the most recent time point in the past is used in the analysis for synchronous data. This method is referred to as LVCF in Table 2. Specifically, for ith subject, at time t_{ij}, if the covariate process $Z_i(\cdot)$ does not have any observation, then the most recently observed $Z_i(s)$ is used, where $s = \max\{x \le t_{ij}, x \in \{s_{i1}, \ldots, s_{iM_i}\}\}$. After this imputation, we proceed with the usual least square estimation procedure.
2. Apply the two-step approach, but use LVCF to estimate γ in the second step. Specifically, for ith subject, in the first stage, obtain the longitudinal residual, $\hat{\omega}_i(t_{ij}) = Y_i(t_{ij}) - X_i(t_{ij})^T\hat{\beta}_c$. In the second stage, regress $\hat{\omega}_i(t_{ij})$ with $Z_i(s)$, where $s = \max\{x \le t_{ij}, x \in \{s_{i1}, \ldots, s_{iM_i}\}\}$. This method is referred to as Centering + LVCF in Table 2.
3. Solve the estimating Eq. (3.17) to obtain estimators of β and γ simultaneously. The estimation of γ is similar to the proposed two-step approach, yet the estimation of β is less efficient than the proposed two-step approach. This method is referred as KS in Table 2.

Table 3
Simulation results of $\hat{\beta}_1$ based on 1000 replications with $h = n^{-0.6}$.

n	PLM				Centering			
	Bias	SD	SE	CP	Bias	SD	SE	CP
$E\{X_2(t)\} = 0.5 + t$								
100	−0.003	0.121	0.116	93	−0.003	0.121	0.116	93
400	−0.007	0.060	0.060	95	−0.007	0.060	0.060	95
900	−0.004	0.040	0.040	95	−0.004	0.040	0.040	95
$E\{X_2(t)\} = 0.5 + t^2$								
100	−0.010	0.125	0.117	92	−0.010	0.125	0.117	93
400	−0.004	0.061	0.060	95	−0.004	0.061	0.060	95
900	−0.005	0.041	0.040	94	−0.005	0.041	0.040	94
$E\{X_2(t)\} = 2sin(2\pi t)$								
100	−0.011	0.122	0.116	93	−0.011	0.122	0.116	93
400	−0.007	0.061	0.060	95	−0.007	0.061	0.060	95
900	−0.007	0.042	0.040	93	−0.007	0.042	0.040	93

Note: "Bias" is the empirical bias, "SD" is the sample standard deviation, "SE" is the average of the standard error estimates, "CP"/100 represents the coverage probability of the 95% confidence interval of $\hat{\beta}_1$. PLM denotes the partial linear model-based method and Centering denotes the centering method.

Automatic bandwidth selection procedure proposed in Section 3.3 is used where the bandwidths are selected in the range $(n^{-0.8}, n^{-0.6})$.

We summarize simulation results in Table 2. For estimation of β, when $E\{Z(t)\} = 2$, all methods perform satisfactorily as assumptions for LVCF are satisfied in this case. When $E\{Z(t)\}$ is non-constant, the performance of LVCF and Centering + LVCF deteriorates and Centering + KS and KS both produce valid results. As our theory predicts, Centering + KS is more efficient than KS for estimation of β, which is reflected on the smaller variance. For estimation of γ and α, both LVCF and Centering + LVCF are biased while Centering + KS and KS produce similar results.

4.3. Correlated longitudinal covariates

In this section, we conduct simulation studies under the relaxed condition ($C2^*$) with the general approach proposed in Section 2.4. The model we use to generate data is

$$Y(t) = \alpha_0 + X_1(t)\beta_1 + X_2(t)\beta_2 + Z(t)\gamma_0 + \epsilon(t), \tag{4.18}$$

where $\alpha_0 = 1, \beta_1 = \beta_2 = 2$ and $\gamma_0 = -1$. We generate 1000 datasets, each consisting of $n = 100, 400$ or 900 subjects. Bandwidth is fixed at $n^{-0.6}$. Other specifications are as follows.

1. Generate a Gaussian processes $X_2(t)$ with mean $0.5 + t, 0.5 + t^2$, or $2sin(2\pi t)$ and $\text{Cov}\{X_2(t), X_2(s)\} = e^{-|t-s|}$.
2. Generate independent Gaussian processes $v_1(t)$ and $v_2(t)$, with $E\{v_1(t)\} = \sqrt{t}, E\{v_2(t)\} = t$, and

$$\text{Cov}\{v_1(t), v_1(s)\} = \text{Cov}\{v_2(t), v_2(s)\} = e^{-|t-s|}.$$

3. The observed covariate processes are constructed as: $X_1(t) = X_2(t) + v_1(t)$ and $Z(t) = X_2(t) + v_2(t)$.
4. Generate a Gaussian process $\epsilon(t)$ with mean zero and

$$\text{Cov}\{\epsilon(t), \epsilon(s)\} = 2^{-|t-s|}.$$

The simulation results are summarized in Table 3. We observe that the partial linear model-based and centering-based methods have almost identical performance. The estimators for β_1 are unbiased. The empirical and model-based standard errors tend to agree, and the coverage probabilities are close to the nominal 95% level. The performance improves with larger sample sizes.

5. Application to the ADNI data

In this section, we illustrate the proposed method of analyzing mixed synchronous and asynchronous longitudinally observed functional data on a study of Alzheimer's disease. In the dataset, 256 subjects were followed for 5 years. The dataset is collected from ADNI GO and ADNI2 in the ADNI study. Among many goals of the ADNI study, we are interested in clinical, functional neuroimaging, and structural variables that affect the progression of mild cognitive impairment and early Alzheimer's disease. The response variable MMSE ranges from 0 to 30, measuring global cognitive performance, where larger values mean a better cognitive state. It is examined from 1 to 7 time points. Baseline covariates include age, years of education, whether the person has a mild cognitive impairment (MCI), whether the person has early Alzheimer's disease (AD), the number of APOE4 gene copies, and the log hazard function of fractional anisotropy (FA) at grid point 0.65. The FA is one of the most popular diffusion-weighted imaging measures that reflects fiber density and myelination in white matter, observed at 1 to 8 time points. Details of the data processing

Table 4
Regression of FA on six covariates.

	Age	Education	MCI	AD	APOE4(1)	APOE4(2)
Fit separately						
Estimate	0.006	0.003	0.057	−0.020	−0.006	−0.070
p-value	0.067	0.634	0.120	0.601	0.875	0.286
All in one model						
Estimate	0.005	0.004	0.075	0.026	−0.013	−0.067
p-value	0.098	0.520	0.101	0.623	0.766	0.326

Table 5
Analysis of dataset from ADNI.

	Naïve		LVCF		Centering+LVCF		Centering+KS		Cao et al. (2015)	
	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
Age	−0.199	0.258	−0.241	0.199	−0.280	0.728	−0.280	0.728	−0.176	0.320
Edu	0.117	0.001	0.111	0.011	0.116	0.000	0.116	0.000	0.106	0.001
MCI	−0.393	0.000	−0.427	0.000	−0.422	0.000	−0.422	0.000	−0.352	0.000
AD	−1.779	0.000	−1.764	0.000	−1.858	0.000	−1.858	0.000	−1.742	0.000
AP4(1)	−0.329	0.000	−0.328	0.000	−0.305	0.000	−0.305	0.000	−0.273	0.000
AP4(2)	−0.391	0.022	−0.366	0.036	−0.415	0.017	−0.415	0.017	−0.296	0.039
FA	(omitted)		0.047	0.225	0.044	0.248	0.061	0.047	0.058	0.062

Note: "AP4(1)" is APOE4(1), and "AP4(2)" is APOE4(2).

are given in Li et al. (2022). The measurement time points of the log hazard functions of FA and the MMSE scores are different between and within subjects. In contrast, baseline measurements align with the MMSE score, giving rise to mixed synchronous and asynchronous longitudinal covariates.

We use model (2.1) to fit the data. Our modeling assumes that the asynchronous longitudinal covariate FA is not correlated with baseline covariates. To better understand this, we regress FA against age, years of education, MCI, AD, APOE4(1), and APOE(2), first in six univariate regression models and then in one multiple regression model. The results are summarized in Table 4. The *p*-values are computed based on the two-sided test. We observe that none of the six baseline covariates are statistically significantly associated with FA. Consequently, we include them all in our model.

We fit model (2.1) with bandwidth chosen in the range $(2(Q_3 - Q_1)n^{-0.7}, 2(Q_3 - Q_1)n^{-0.6})$, where Q_3 and Q_1 are third and first quantiles of the combined observation times of MMSE and FA, $n = 256$ is number of patients after eliminating missing data (Little and Rubin, 2014). We implemented five methods: Naïve, LVCF, Centering + LVCF, Centering + KS, and KS. The naïve method ignores FA in fitting the linear regression model. In the two-step approach, regression coefficients of age, education, MCI, AD, APOE4(1), and APOE4(2) are estimated in the first step using the centering approach. The regression coefficient of FA is estimated in the second step using either LVCF (Centering + LVCF) or kernel smoothing (Centering + KS). LVCF and KS estimate all regression coefficients simultaneously. We normalize continuous variables, including MMSE, age, education, and HA. Analysis results are summarized in Table 5.

There is no statistically significant association between age and MMSE across all methods. As education, MCI, AD, APOE(1), and APOE(2) are baseline covariates, their parameter estimates are similar across different methods, and all show statistical significance. The estimation results show that MCI at baseline, AD at baseline, APOE4(1), and APOE4(2) have significant negative effects on MMSE. In contrast, education plays a positive role, which has been verified in the literature (Bekris et al., 2010). For misaligned FA, LVCF and Centering + LVCF do not show statistical significance. However, the newly proposed Centering + KS shows a statistically significant positive effect. This finding is consistent with the existing literature that lower FA values, which means less white matter, are associated with lower MMSE scores (Kristensen et al., 2019).

6. Concluding remarks

In this paper, we propose valid statistical approaches for analyzing longitudinal data with omitted longitudinal covariates. Furthermore, to deal with mixed synchronous and asynchronous longitudinal covariates, we propose a two-step approach for analysis. In the first step, a partial linear model or a centering approach is used to estimate the regression coefficient of the synchronous longitudinal covariate. In the second step, we regress the asynchronous longitudinal covariate with longitudinal residual from the first step through kernel weighting to obtain regression coefficient estimation of the asynchronous longitudinal covariate. We obtain parametric root n rate of convergence for regression coefficient estimation of the synchronous longitudinal covariate and $n^{2/5}$ rate of convergence for regression coefficient estimation of the asynchronous longitudinal covariate. In terms of implementation, the centering approach is computationally faster, while the partial linear model approach can suggest possible forms of the omitted longitudinal covariate. We require the unconfoundedness assumption specified in ($C2^*$) on the covariates, which is plausible in randomized clinical trials. In observational studies, this assumption is very unlikely to hold. We suggest modeling the synchronous and asynchronous longitudinal covariates jointly, as presented in Section 3.2, to get unbiased regression coefficient estimation.

Per the request of a referee, for sensitivity analysis, we conducted simulation studies under the same set-up of Section 4.3 where condition (C2) is not satisfied and implemented the strategy of Section 2.3. The results are relegated in the Supplementary Material.

We use a working independence covariance matrix for analysis. Pepe and Anderson (1994) pointed out that a working independence covariance matrix is a safe choice to get an unbiased estimation of the regression coefficient in longitudinal data analysis with time-varying covariates. A carefully chosen working covariance matrix can lead to efficiency gain compared to the simple independent covariance matrix. A fully efficient estimator requires a correctly specified working covariance matrix, which is difficult in practice. This is similar in spirit to using ordinary least squares in the presence of heteroscedasticity in linear models.

Acknowledgments

This research is partially supported by the National Science Foundation of China [No. 12171483]. We thank Ting Li for help with data acquisition, Peng Ding for helpful discussions, and Congmin Liu for help in the data analysis. Data used to prepare this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jspi.2023.106135. Supplementary Material includes detailed proofs of theoretical results and additional simulations.

References

Bates, S., Kennedy, E., Tibshirani, R., Ventura, V., Wasserman, L., 2022. Causal inference with orthogonalized regression adjustment: taming the phantom. arXiv:2201.13451v2.

Bekris, L.M., Yu, C.-E., Bird, T.D., Tsuang, D.W., 2010. Genetics of alzheimer's disease. J. Geriatr. Psychiatry Neurol. 23, 213–227.

Cao, H., Li, J., Fine, J.P., 2016. On last observation carried forward and asynchronous longitudinal regression analysis. Electron. J. Stat. 10, 1155–1180.

Cao, H., Zeng, D., Fine, J.P., 2015. Regression analysis of sparse asynchronous longitudinal data. J. R. Stat. Soc. Ser. B Stat. Methodol. 77, 755–776.

Chen, L., Cao, H., 2017. Analysis of asynchronous longitudinal data with partially linear models. Electron. J. Stat. 11, 1549–1569.

Ding, P., 2021. The Frisch-Waugh-Lovell theorem for standard errors. Statist. Probab. Lett. 168, 108945.

Fan, J., Li, R., 2004. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Amer. Statist. Assoc. 99, 710–723.

Frisch, R., Waugh, F.V., 1933. Partial time regressions as compared with individual trends. Econometrica 1, 387–401.

Kristensen, T.D., Mandl, R.C.W., Raghava, J.M., Jessen, K., Jepsen, J.R.M., Fagerlund, B., Glenthøj, L.B., Wenneberg, C., Krakauer, K., Pantelis, C., Nordentoft, M., Glenthøj, B.Y., Ebdrup, B.H., 2019. Widespread higher fractional anisotropy associates to better cognitive functions in individuals at ultra-high risk for psychosis. Hum. Brain Mapp. 40, 5185–5201.

Li, T., Li, T., Zhu, Z., Zhu, H., 2022. Regression analysis of asynchronous longitudinal functional and scalar data. J. Amer. Statist. Assoc. in press.

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

Lin, D.Y., Ying, Z., 2001. Semiparametric and nonparametric regression analysis of longitudinal data. J. Amer. Statist. Assoc. 96, 103–126.

Little, R.J.A., Rubin, D.B., 2014. Statistical Analysis with Missing Data. John Wiley & Sons.

Lovell, M.C., 1963. Seasonal adjustment of economic time series and multiple regression analysis. J. Amer. Statist. Assoc. 58, 993–1010.

Lovell, M.C., 2008. A simple proof of the FWL theorem. J. Econ. Educ. 39, 88–91.

Nadaraya, E.A., 1964. On estimating regression. Theory Probab. Appl. 9, 141–142.

Pepe, M.S., Anderson, G.L., 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Comm. Statist. Simulation Comput. 23, 939–951.

Qian, L., Wang, S., 2017. Subject-wise empirical likelihood inference in partial linear models for longitudinal data. Comput. Statist. Data Anal. 111, 77–87.

Sentürk, D., Dalrymple, L.S., Mohammed, S.M., Kaysen, G.A., Nguyen, D.V., 2013. Modeling time-varying effects with generalized and unsynchronized longitudinal data. Stat. Med. 32, 2971–2987.

Sun, D., Zhao, H., Sun, J., 2021. Regression analysis of asynchronous longitudinal data with informative observation processes. Comput. Statist. Data Anal. 107161.

Watson, G.S., 1964. Smooth regression analysis. Sankhya: Indian J. Stat. Ser. A 26, 359–372.

Xiong, X., Dubin, J.A., 2010. A binning method for analyzing mixed longitudinal data measured at distinct time points. Stat. Med. 29, 1919–1931.