Journal of Statistical Planning and Inference 139 (2009) 2933 – 2951

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

On locally weighted estimation and hypothesis testing of varying-coefficient models with missing covariates

Heung Wong[a], Shaojun Guo[b,*], Min Chen[b], Wai-Cheung IP[a]

[a]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*
[b]*Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China*

ARTICLE INFO

Article history:
Received 17 July 2007
Received in revised form
8 January 2009
Accepted 29 January 2009
Available online 7 February 2009

Keywords:
Varying-coefficient models
Local linear smoother
Locally weighted estimating equation
Missing at random

ABSTRACT

Varying-coefficient model $Y=\sum_{j=1}^{p}\beta_j(U)X_j+\varepsilon$ has been studied extensively when data are completely observed. When the covariates X are missing at random, we propose a locally weighted estimator based on the inverse selection probabilities. Distribution theory of $\hat{\beta}(\cdot)$ is derived when the selection probabilities are known, estimated parametrically or nonparametrically. We show that the resulting nonparametric estimator of $\hat{\beta}(\cdot)$ when the selection probabilities are estimated nonparametrically has a smaller asymptotic variance than that when the selection probabilities are known or estimated parametrically. Motivated by Robin et al. [1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866], we also consider simple locally augmented weighted estimator. However, we show that it does not improve the efficiency theoretically. We have constructed a bootstrap test for goodness of fit of models in the missing covariates case. The results of a simulation study are also given to illustrate our method. The proposed method is applied to analyze an AIDS dataset from a clinical study.

1. Introduction

The analysis of data with missing covariates is a practical and interesting problem in regression analysis. Developing methods for regression analysis with missing covariates have been an active research area in the past decade (see Little, 1992; Little and Rubin, 1987 for a review). Parametric regression models with missing data have been widely used for analyzing real datasets and for providing a parsimonious description of the relationship between the response variable and its covariates. For example, we often choose a linear regression $Y = X^{\tau}\beta + \varepsilon$ or a nonlinear parametric regression $Y = g(X, \beta) + \varepsilon$ for a specific function g to investigate the dependence of a continuous response variable on a vector of covariates (for example, see Robins et al., 1994). However, parametric models have the risk of introducing modelling biases. To relax the assumptions on classical parametric forms, nonparametric and semiparametric regression models have been proposed, which provide more useful insights. In dealing with nonparametric regression problems with multiple covariates, many powerful approaches have been proposed to avoid the 'curse of dimensionality', such as additive models (Hastie and Tibishirani, 1990), low-dimensional interaction models (Friedman, 1991; Gu and Wahba, 1993), multiple-index models (Härdle and Stoker, 1989; Li, 1991), partially linear models (Wahba, 1984; Green and Silverman, 1994), and their hybrids (Carroll et al., 1997; Fan et al., 1998; Heckman et al., 1998), among others. An important alternative to these powerful approaches is the varying-coefficient models (Cleveland et al., 1991; Hastie and Tibishirani, 1993;

* Corresponding author.
E-mail address: guoshaoj@amss.ac.cn (S. Guo).

0378-3758/$ - see front matter Crown Copyright © 2009 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.01.016

Fan and Zhang, 1999, 2000a, b). The varying-coefficient model is as follows:

$$Y = \sum_{j=1}^{p} \beta_j(U)X_j + \varepsilon, \tag{1.1}$$

for given covariates $(U, X_1, \ldots, X_p)^\tau$ and response variable Y with

$$E\{\varepsilon|U, X_1, \ldots, X_p\} = 0,$$

and

$$var\{\varepsilon|U, X_1, \ldots, X_p\} = \sigma^2(U).$$

If $\beta_j(u_0) \equiv \beta_j$, $j = 1, \ldots, p$, model (1.1) becomes the classical linear model, and if $X_1 \equiv 1$ and $\beta_j(u_0) \equiv \beta_j$, $j = 2, \ldots, p$, model (1.1) becomes the partially linear model.

Model (1.1) is very attractive because it has a meaningful interpretation and still retains general nonparametric characteristics. There is, however, little literature on missing covariates in varying-coefficient models. In this article, we consider the problem of estimation and testing of the varying-coefficient model when covariates X are not fully observed for some study subjects by design or by happenstance. A good example of this situation is, in Section 6, an AIDS clinical trial group (ACTG 315) study, which investigated the relationship between virologic and immunologic responses in AIDS clinical trials. In this study, some CD4+ cell counts are missing because the covariate and the viral load were measured at different times. In general, it is believed that the virologic response (measured by viral load) and immunologic response (measured by CD4+ cell counts) are negatively correlated during antiviral treatments. However, their relationship may not be constant during the whole period of treatment.

Several kinds of estimation and testing methods have been developed in Model (1.1) when the data are completely observed. Examples include penalized least squares (Wahba, 1984; Chiang et al., 2001), kernel smoothing (Hoover et al., 1998; Wu et al., 1998, 2000), local polynomial approach (Fan and Zhang, 1999, 2000a, b; Zhang et al., 2000), series approximation (Huang et al., 2002, 2004), smoothing spline (Eubank et al., 2004) and so on.

A 'naive' way of handling incomplete data is a complete data analysis, which uses only the subjects with complete covariates values. This method is usually called thecomplete-case method. The complete-case method may not only lose efficiency due to discarding incomplete observations, but also may generate biased estimators when the missing-data mechanism depends on outcome variables. Tsiatis (2006) explained this problem in detail in Chapter 6 and Liang et al. (2004) gave some bias calculations for partially linear models with missing covariates in Section 5. We also make some direct bias calculations in Section 5. These all motivate us to make some modifications to the complete-case method. To gain efficiency, when missing mechanism is missing at random (MAR) in the sense of Rubin (1976), Flander and Greenland (1991), and Zhao and Lipsitz (1992) suggested a simple weighted estimator based on weighted estimating equation for parametric analysis. Wang et al. (1998) gave a local version for generalized linear model with missing covariates, based on the inverse probability-weighted idea of Horvitz and Thompson (1952). Wang et al. (1997, 1998) called these selection probabilities. In this article, we propose to use this locally weighted estimating equation (LWEE) technique to handle our problem. In our problem, we propose to estimate the selection probabilities by nonparametric kernel smoothers, although it may be prespecified at design stage in some applications. Of course, we may model the missing data probabilities via a logistic regression with a vector of nuisance parameters to be estimated from data when the selection probabilities is correctly specified. However, misspecification may lead to biased estimates for the regression parameters.

For parametric regression analysis, the Horvitz–Thompson (H–T) weighting scheme has a curious and important property. Consider two estimators: (a) one with known selection probabilities and weights and (b) one where the selection probabilities are estimated by a properly specified parametric model or a nonparametric model. The two methods both yield consistent estimates, but that with estimated weights generally has a smaller asymptotic variance (Robins et al., 1994). A heuristic argument of this phenomenon was given in Robins et al. (1994, Section 6.1), and a more detailed illustration has been studied by Wang et al. (1997). Wang et al. (1998) expected the same sort of results to hold in the nonparametric regression case with nonparametrically estimated selection probabilities. However, they showed that whether weights were estimated or not had no effect on asymptotic variance, while it did have an effect on bias in general. Fortunately, we show that the above parametric H–T property continues to hold when we estimate thecoefficient functions $\{\beta_j(u), j = 1, \ldots, p\}$ with the selection probabilities estimated nonparametrically. This is because our model (1.1) has the special quasi-linear structure and the index variable u is always observed.

To gain efficiency, augmented estimating equations were proposed by Robins et al. (1994) for semiparametric models and by Robins et al. (1995) for parametric models. This new class of estimators has two attractive advantages. One is double robust property. That is, such estimating equations yield consistent regression parameter estimators when either the model for the selection probabilities or the model for the covariates is correctly specified (see Van der Laan and Rubins, 2003; Tsiatis, 2006). Secondly, Robins et al. (1994) showed that the optimal estimator in their new class attained the semiparametric variance bound. However, the efficient score usually does not have a closed form, and it requires solving a functional integral equation. Following this idea, Liang et al. (2004) applied this technique to semiparametric partially linear models and obtained a relatively efficient estimator of the parametric part. They gave the closed form of the parametric estimator, which could be calculated easily and stably. Motivated by these, we also try to construct the corresponding simple locally augmented weighted estimating equation (SLAWEE). (For details see Remark 1 of Section 2 in this article.) However, in practice, we realize that the corresponding computing

program is not often stable, although it has good properties theoretically. On the other hand, we show that the asymptotic variance of the estimator based on SLAWEE is the same as that based on LWEE with selection probabilities estimated nonparametrically. Therefore, the estimator we propose is reasonable.

Another important statistical question in fitting model (1.1) is that if there exists a parametric structure for $\beta_j(\cdot)$ for $j=1,\ldots,p$. A testing procedure is proposed based on the comparison of the sum of the inverse weighted residual squares under null and alternative models which is an extension of the generalized likelihood ratio tests proposed by Fan et al. (2001) to the case with incomplete data. A wild bootstrap method is proposed for finding the null distribution of the test statistic. Our simulation study shows the resulting testing procedure is indeed powerful and the bootstrap method does give the right null distribution.

This article is organized as follows. Section 2 describes the model and estimating methodology. In Section 3, we present the asymptotic properties. A wild bootstrap-based test is proposed in Section 4. Section 5 examines some finite sample properties of the proposed estimators by applying them to some simulated datasets and Section 6 analyzes the real data from an ACTG study of viral load. Section 7 provides further discussions. All the detailed proofs are given in the Appendix.

2. Models and methodology

2.1. The models

Let $(Y_1,\delta_1X_1,U_1,\delta_1),(Y_2,\delta_2X_2,U_2,\delta_2),\ldots,(Y_n,\delta_nX_n,U_n,\delta_n)$ be a set of independent random variables from model (1) where, for each i, $\delta_i=1$ if X_i is observed and $\delta_i=0$ otherwise, Y_i,U_i are always observed and U_i has a density function $f(\cdot)$ bounded away from 0. In this article, we assume the X's are MAR in the sense that

$$P(\delta_i=1|Y_i,X_i,U_i)=P(\delta_i=1|Y_i,U_i)=\pi(Y_i,U_i)>0. \tag{2.1}$$

We first assume that the selection probability $\pi(Y,U)$ is known. And, in Section 2.3, we discuss the estimation of selection probability $\pi(Y,U)$ when it is unknown.

When X is also observable completely, we apply a local linear technique to estimate the varying-coefficient function $\beta_j(\cdot)$ $(j=1,\ldots,p)$ about model (1.1). For each given point u_0, approximate the function locally as

$$\beta_j(u)\approx a_j+b_j(u-u_0),$$

for u in a neighborhood of u_0. This leads to the following weighted local least-square estimator problem:

minimizing

$$\sum_{i=1}^{n}\left[Y_i-\sum_{j=1}^{p}\{a_j+b_j(U_i-u_0)\}X_{ij}\right]^2K_h(U_i-u_0), \tag{2.2}$$

for a given kernel function $K(\cdot)$ and bandwidth h, where $K_h(\cdot)=K(\cdot/h)/h$ (see also Fan and Zhang, 1999; Cai et al., 2000 for details). Denote $\tilde{X}_i=(X_{i1},X_{i1}(U_i-u_0)),\ldots,X_{ip},X_{ip}(U_i-u_0))$, then the solution to the local least-square problem (2.2) can easily be obtained and is equivalent to solve the following local estimating equation:

$$\sum_{i=1}^{n}K_h(U_i-u_0)\tilde{X}_i^{\tau}\left[Y_i-\sum_{j=1}^{p}\{a_j+b_j(U_i-u_0)\}X_{ij}\right]=0. \tag{2.3}$$

About local estimating equation, see Carroll et al. (1998) for details. Especially, the above local estimating equation (2.3) can be solved in closed form. Now, we extend this technique of local estimating equation to handle the case with missing covariates.

2.2. Estimation based on locally weighted estimating equation (LWEE)

In the spirit of Horvitz and Thompson (1952), we construct the following LWEE:

$$\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(U_i-u_0)\tilde{X}_i^{\tau}\left[Y_i-\sum_{j=1}^{p}\{a_j+b_j(U_i-u_0)\}X_{ij}\right]=0, \tag{2.4}$$

where the weight $\{\pi_i=\pi(Y_i,U_i)\}_{i=1}^n$ is defined as (2.1) and $K_h(\cdot)$ is a kernel function and h is a bandwidth. See also Wang et al. (1997, 1998). It is easy to see that when the data are complete, the solution to (2.4) is equivalent to local linear estimator of regression coefficient. Denote $e_{k,m}$ the unit vector of length m with one at the k-th position. Denote $\hat{\beta}_{W,k}(u_0,\pi),\hat{b}_{W,k}(u_0,\pi)$ $(k=1,\ldots,p)$ the solution to the above LWEE (2.4). If $\sum_{i=1}^{n}(\delta_i/\pi_i)K_h(U_i-u_0)\tilde{X}_i^{\tau}\tilde{X}_i$ is invertible, we can obtain explicitly that

$$\hat{\beta}_{W,k}(u_0,\pi)=e_{2k-1,2p}^{\tau}\left[\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(U_i-u_0)\tilde{X}_i^{\tau}\tilde{X}_i\right]^{-1}\left[\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K_h(U_i-u_0)\tilde{X}_i^{\tau}Y_i\right]. \tag{2.5}$$

Remark 1. For general parametric models $E\{Y|X\} = g(X,\theta)$, Robins et al. (1994) proposed the following augmented weighted estimating equation:

$$\psi(\cdot,\theta) = \frac{\delta}{\pi}p(X)\{Y - g(X,\theta)\} - \frac{\delta-\pi}{\pi}\phi(Y,\theta), \tag{2.6}$$

for some user-supplied function $p(x)$, where $\phi(Y,\theta)$ is a general function and the optimal choice of $\phi(Y,\theta)$ is $\phi(Y,\theta) = E\{p(X)(Y - g(X,\theta))|Y\}$. See also Liang et al. (2004). Motivated by the above idea, we also construct the corresponding SLAWEE:

$$\sum_{i=1}^{n} K_h(U_i - u_0)\left\{\frac{\delta_i}{\pi_i}\tilde{X}_i^{\tau}\left[Y_i - \sum_{j=1}^{p}(a_j + b_j(U_i - u_0))X_{ij}\right] - \frac{\delta_i - \pi_i}{\pi_i}\phi(Y_i,U_i,\theta)\right\} = 0, \tag{2.7}$$

where

$$\phi(y,u,\theta) = E\left\{\tilde{X}^{\tau}\left[Y - \sum_{j=1}^{p}(a_j + b_j(U - u_0))X_j\right]\middle| y,u\right\}, \tag{2.8}$$

$\theta = (a_1,b_1,\ldots,a_p,b_p)$, $\pi_i = \pi(Y_i,U_i)$, and $K_h(\cdot)$ is kernel function and h is a bandwidth.

Denote $\hat{\beta}_{A,k}(u_0,\pi), \hat{b}_{A,k}(u_0,\pi)(k = 1,\ldots,p)$ the solution to the above SLAWEE (2.7). Denote $\hat{\phi}(y,u,\theta) = \hat{E}^{y,u}(\tilde{X}^{\tau})y - \hat{E}^{y,u}(\tilde{X}^{\tau}\tilde{X})\theta$, where $\hat{E}^{y,u}(\tilde{X}^{\tau})$ and $\hat{E}^{y,u}(\tilde{X}^{\tau}\tilde{X})$ are the H–T bivariate local linear estimators of $E(\tilde{X}^{\tau}|y,u)$ and $E(\tilde{X}^{\tau}\tilde{X}|y,u)$, see (2.10).

Note that $\hat{\phi}(Y,U,\theta)$ is an estimator of $\phi(Y,U,\theta)$ and linear in θ. We substitute $\hat{\phi}(Y,U,\theta)$ into (2.7) and obtain that

$$\hat{\beta}_{A,k}(u_0,\pi) = e_{2k-1,2p}^{\tau}\left[\sum_{i=1}^{n} K_h(U_i - u_0)\left(\frac{\delta_i}{\pi_i}\tilde{X}_i^{\tau}\tilde{X}_i - \frac{\delta_i - \pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}\tilde{X})\right)\right]^{-1}\left[\sum_{i=1}^{n} K_h(U_i - u_0)\left(\frac{\delta_i}{\pi_i}\tilde{X}_i^{\tau} - \frac{\delta_i - \pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau})Y_i\right)\right]. \tag{2.9}$$

Here, estimating missing covariates $\hat{E}^{y,u}(\tilde{X})$ and $\hat{E}^{y,u}(\tilde{X}^{\tau}\tilde{X})$ is the important point. We can estimate them using the H–T bivariate local linear estimation with the complete-case data. In detail, the H–T bivariate local linear estimator of the j-th element $\hat{E}^{y,u}(X_j)$ is the solution of α_0 to

$$\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}K^{\star}_{\lambda_1,\lambda_2}(Y_i - y, U_i - u)\begin{pmatrix}1\\ Y_i - y\\ U_i - u\end{pmatrix}\{X_{ij} - \alpha_0 - \alpha_1(Y_i - y) - \alpha_2(U_i - u)\} = 0, \tag{2.10}$$

where the weight $\{\pi_i = \pi(Y_i,U_i)\}_{i=1}^{n}$ is defined as in (2.1), $K^{\star}(\cdot,\cdot)$ is a two-dimensional density function with bandwidths λ_1 and λ_2 and X_{ij} is the j-th element of X_i. In general, the choice of $K^{\star}(y,u)$ is $K(y)K(u)$ where $K(\cdot)$ is a symmetric density function with a compact support. As discussed in the Introduction, in practice, this local estimator needs more bandwidth selections and is often unstable for finite sample, although it has relatively small asymptotic variance.

Remark 2. As discussed in Fan and Zhang (1999), a one-step procedure is not optimal when coefficient functions admit different degrees of smoothness. To achieve the optimal rate, a two-step procedure can be extended to the missing covariate case. we assume that β_p possesses a bounded fourth derivative so that the function can locally be approximated by a cubic function,

$$\beta_p(u_0) \approx a_p + b_p(u - u_0) + c_p(u - u_0)^2 + d_p(u - u_0)^3,$$

for u in a neighborhood of u_0. In the first step, we can get an initial estimate of $\beta_1(\cdot),\ldots,\beta_{p-1}(\cdot)$ using local linear technique for a given initial bandwidth h_1 and kernel function $K(\cdot)$. Then, in the second step, we substitute the preliminary estimates $\hat{\beta}_{1,0}(\cdot),\ldots,\hat{\beta}_{p-1,0}(\cdot)$ and use a local cubic fit to estimate $\beta_p(u_0)$, namely, solve the following LWEE:

$$\sum_{i=1}^{n} K_{h_2}(U_i - u_0)\frac{\delta_i}{\pi_i}\tilde{X}_{ip}^{\tau}\left\{Y_i - \sum_{j=1}^{p-1}\hat{\beta}_{j,0}(U_i)X_{ij} - \tilde{X}_{ip}\theta_p\right\} = 0,$$

where

$$\tilde{X}_p = (X_p, X_p(U - u_0), X_p(U - u_0)^2, X_p(U - u_0)^3),$$
$$\theta_p = (a_p, b_p, c_p, d_p)^{\tau},$$

and h_2 is the bandwidth in the second step.

Denote the second step estimator of $\beta_p(u_0)$ by $\hat{\beta}_{TS,p}(u_0)$. Similar to the proof of Theorem 2 of Fan and Zhang (1999), we can prove the asymptotic properties of $\hat{\beta}_{TS,p}(u_0)$, which are similar to the results in Theorem 2 of Fan and Zhang (1999).

2.3. Estimation of selection probabilities

In Eq. (2.4), the missing probabilities π_i $(i = 1, \ldots, n)$ are generally unknown parameters (see also Wang et al., 1997). One simple approach is to assume a parametric function of π_i, as also discussed by Robins et al. (1994). For instance, we suppose

$$\pi_i = \pi(Y_i, U_i) = \{1 + \exp(-\alpha_0 - \alpha_1 Y_i - \alpha_2 U_i)\}^{-1} = \pi_i(\alpha), \tag{2.11}$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2)^\tau$ are unknown vector parameters. (See le Cessie and van Houwelingen, 1991 for a global test statistic using this model assumption.) We can use the maximum likelihood method to estimate α, denoted by $\hat{\alpha}$. Denote locally weighted estimator with $\pi(\hat{\alpha})$ by $\hat{\beta}_{WP,k}(u_0, \hat{\pi})$. Its asymptotic distribution is given in Theorem 2 when the model for $\pi(Y, U)$ is correctly specified as (2.11).

Nonparametric smoothing is also a useful tool for estimating the selection probabilities. Copas (1983) proposed kernel estimates to plot a binary response against covariates. For convenience, denote $W = (Y, U)$ and $w = (y, u)$. Let $L(\cdot)$ be a two-dimensional symmetric density function. Thus, based on the data $\{(\delta_i, W_i), i = 1, \ldots, n\}$, the kernel smoother of $\pi(w)$ can be given by

$$\hat{\pi}(w) = \frac{\sum_{j=1}^n \delta_j L_{h_0}(w - W_j)}{\sum_{j=1}^n L_{h_0}(w - W_j)}, \tag{2.12}$$

where $L_h(\cdot) = L(\cdot/h)/h^2$ and h_0 is a bandwidth. As we know, kernel smoothing depends on the choice of the bandwidth. An important problem associated with the use of the estimator is the selection of a good value for the bandwidth (see Eubank, 1988; Fan and Gijbels, 1996 for more details).

Denote locally weighted estimator with estimated selection probabilities (2.12) by $\hat{\beta}_{WS,k}(u_0, \hat{\pi})$. Intuitively, using more information in the estimation may improve efficiency. In fact, $\hat{\beta}_{WS,k}(u_0, \hat{\pi})$ gets a smaller asymptotic variance than $\hat{\beta}_{WS,k}(u_0, \pi)$. Its asymptotic distribution is given in Theorem 3.

3. Asymptotic properties

3.1. Main asymptotic results

For convenience, we make the following notation. Denote $AA^\tau = A^{\otimes 2}$, $\mu_i = \int t^i K(t)\,dt$ and $\nu_i = \int t^i K^2(t)\,dt$, $i = 1, 2$. Denote $r_{ij} = r_{ij}(u_0) = E(X_i X_j | U = u_0)$ for $i = 1, \ldots, p$. Put

$$\begin{aligned}
&\alpha_j(u) = (r_{1j}(u), \ldots, r_{pj}(u))^\tau, \quad j = 1, \ldots, p,\\
&\Omega_i(u) = E\{(X_1, \ldots, X_i)^\tau (X_1, \ldots, X_i)^\tau | U = u\},\\
&\Omega_i = \Omega_i(u_0), \quad i = 1, \ldots, p.
\end{aligned}$$

Theorem 1. *Suppose $\pi(w)$ is known. Under the assumptions* (a)–(g) *in the Appendix, then $\hat{\beta}_{W,k}(u_0, \pi)$ $(k = 1, \ldots, p)$ has the property*

$$\sqrt{nh}\left(\hat{\beta}_{W,k}(u_0, \pi) - \beta_k(u_0) - \frac{1}{2}h^2 \mu_2 \beta_k''(u_0) + o(h^2)\right) \xrightarrow{\mathscr{D}} N\left(0, \frac{v_0}{f(u_0)} e_{k,p}^\tau \Omega_p^{-1} \Omega_p^\star \Omega_p^{-1} e_{k,p}\right), \tag{3.1}$$

where

$$\Omega_p^\star = E\left\{\frac{\varepsilon^2}{\pi} X^\tau X | U = u_0\right\}.$$

Theorem 2. *Assume that $\pi(w)$ follows the model* (2.11) *with unknown nuisance parameters α. Under the assumptions* (a)–(g) *in the Appendix, then we obtain that, for $k = 1, \ldots, p$,*

$$\sqrt{nh}\left(\hat{\beta}_{WP,k}(u_0, \hat{\pi}) - \beta_k(u_0) - \frac{1}{2}h^2 \mu_2 \beta_k''(u_0) + o(h^2)\right) \xrightarrow{\mathscr{D}} N\left(0, \frac{v_0}{f(u_0)} e_{k,p}^\tau \Omega_p^{-1} \Omega_p^\star \Omega_p^{-1} e_{k,p}\right), \tag{3.2}$$

where

$$\Omega_p^\star = E\left\{\frac{\varepsilon^2}{\pi} X^\tau X | U = u_0\right\}.$$

Theorem 3. *Assume that for some constant c, $\pi(w) \geqslant c > 0$ is a smoothing function of w such that the second derivative of π exists and is continuous. Under the assumptions* (a)–(h) *in the Appendix, then, for $k = 1, \ldots, p$,*

$$\sqrt{nh}\left(\hat{\beta}_{WS,k}(u_0, \hat{\pi}) - \beta_k(u_0) - \frac{1}{2}h^2 \mu_2 \beta_k''(u_0) + o(h^2)\right) \xrightarrow{\mathscr{D}} N\left(0, \frac{v_0}{f(u_0)} e_{k,p}^\tau \Omega_p^{-1} \tilde{\Omega}_p^\star \Omega_p^{-1} e_{k,p}\right), \tag{3.3}$$

where

$$\tilde{\Omega}_p^{\star} = E\left\{\frac{\varepsilon^2}{\pi}X^{\tau}X|U=u_0\right\} - E\left\{\frac{1-\pi}{\pi}[E(X^{\tau}\varepsilon|Y,U)]^{\otimes 2}|U=u_0\right\}.$$

Theorem 4. *Assume that for some constant c, $\pi(w)\geqslant c>0$ is a smoothing function of w such that the second derivative of π exists and is continuous. Under the assumptions* (a)–(f) *and* (h′) *in the Appendix, then, for $j=1,\ldots,p$, $\hat{\beta}_{A,k}(u_0,\pi)$ and $\hat{\beta}_{A,k}(u_0,\hat{\pi})$ have the same limiting distribution:*

$$\sqrt{nh}\left(\hat{\beta}_{A,k}(u_0)-\beta_k(u_0)-\frac{1}{2}h^2\mu_2\beta_k''(u_0)+o(h^2)\right)\xrightarrow{\mathscr{D}} N\left(0,\frac{v_0}{f(u_0)}e_{k,p}^{\tau}\Omega_p^{-1}\tilde{\Omega}_p^{\star}\Omega_p^{-1}e_{k,p}\right), \tag{3.4}$$

where

$$\tilde{\Omega}_p^{\star} = E\left\{\frac{\varepsilon^2}{\pi}X^{\tau}X|U=u_0\right\} - E\left\{\frac{1-\pi}{\pi}[E(X^{\tau}\varepsilon|Y,U)]^{\otimes 2}|U=u_0\right\}.$$

3.2. Comparison with different estimators

From the above asymptotic results, we see that

(1) Under the assumptions (a)–(h′) in the Appendix, all the estimators has the same asymptotic bias.
(2) The asymptotic variance of $\hat{\beta}_{WP,k}(u_0,\hat{\pi})$ is equal to that of $\hat{\beta}_{W,k}(u_0,\pi)$, which implies that there is no gain in efficiency using parametric selection probabilities.
(3) The asymptotic variance of $\hat{\beta}_{WS,k}(u_0,\hat{\pi})$ is less than that of $\hat{\beta}_{W,k}(u_0,\pi)$ and $\hat{\beta}_{WP,k}(u_0,\hat{\pi})$. Therefore, we can get a more efficient estimator using the selection probabilities estimated nonparametrically than even if the selection probabilities are specified correctly.
(4) When π is unknown, $\hat{\beta}_{WS,k}(u_0,\hat{\pi})$ has the same asymptotic normal distribution as $\hat{\beta}_{A,k}(u_0,\pi)$ and $\hat{\beta}_{A,k}(u_0,\hat{\pi})$. That is to say, they are asymptotically equivalent. However, our estimated method is simpler and more stable computationally. Therefore, locally weighted estimation method is a good choice.

4. Bootstrap-based goodness-of-fit test

To test whether model (1.1) reduces to a specified parametric model such as the linear regression models and partially linear regression models, we propose a goodness-of-fit test based on the comparison of the sum of weighted residual squares (WRSS) for both parametric and semiparametric fittings. This method is an extension of the generalized likelihood techniques developed by Fan et al. (2001) to the case of missing covariates.

Consider the null hypothesis:

$$H_0\colon \beta_j(u)=\alpha_j(u,\theta_j),\quad j=1,\ldots,p, \tag{4.1}$$

where $\alpha_j(u,\theta_j)$ is a given family of functions indexed by unknown parameter vector θ_j. Let $\hat{\theta}_j$ be an estimator of θ_j. The sum of weighted residual squares under the null hypothesis is

$$WRSS_0=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\left[Y_i-\sum_{j=1}^{p}\alpha_j(U_i,\hat{\theta}_j)X_{ij}\right]^2.$$

Analogously, the sum of weighted residual squares corresponding to model (1.1) is

$$WRSS_1=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\left[Y_i-\sum_{j=1}^{p}\hat{\beta}_{WS,j}(U_i)X_{ij}\right]^2.$$

Then, the test statistic is defined as

$$T_n=\frac{WRSS_0-WRSS_1}{WRSS_1}=\frac{WRSS_0}{WRSS_1}-1$$

and we reject the null hypothesis (4.1) for large value of T_n. We use the following nonparametric bootstrap approach to evaluate the *p*-value of the test.

Step 1: Suppose the number of complete data is m. Generate the bootstrap residuals $\{\varepsilon_i^\star\}_{i=1}^m$ from the empirical distribution of the centralized residuals $\{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}\}_{i=1}^m$, based on complete data, where

$$\hat{\varepsilon}_i = Y_i - \sum_{j=1}^{p} \hat{\beta}_{WS,j}(U_i)X_{ij}, \quad \bar{\hat{\varepsilon}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\varepsilon}_i, \quad \delta_i = 1,$$

and define

$$Y_i^\star = \begin{cases} \sum_{j=1}^{p} \alpha_j(U_i, \hat{\theta}_j)X_{ij} + \varepsilon_i^\star, & \delta_i = 1, \\ Y_i, & \delta_i = 0. \end{cases}$$

Step 2: Calculate the bootstrap test statistic $T_n^\star$ based on the sample $\{Y_i^\star, U_i, X_i, \delta_i\}_{i=1}^n$.

Step 3: Reject the null hypothesis H_0 when T_n is greater than the upper-α point of the conditional distribution of $T_n^\star$ given $\{Y_i^\star, U_i, X_i, \delta_i\}_{i=1}^n$.

The p-value of the test is simply the relative frequency of the event $\{T_n^\star \geqslant T_n\}$ in the replications of the bootstrap sampling. For the sake of simplicity, we can use the same bandwidth in calculating $T_n^\star$ as that in T_n. Note that we bootstrap the centralized residuals from the nonparametric fit instead of the parametric fit, because the nonparametric estimate of the residual is always consistent, no matter the null or the alternative hypothesis is correct.

5. Simulation studies

In this section, we conduct simulation studies to examine the finite sample performance of our proposed method. Recall that $\hat{\beta}_{CC,k}(\cdot)$ is the complete-case (CC) method which solves (2.3) with complete data only. We use the Epanechnikov kernel $K(u) = 0.75(1-u^2)I(|u| \leqslant 1)$ and kernel $L(w) = K(y)K(u)$ in all simulations. To select the bandwidth h_0 and h, the bandwidth selector proposed by Ruppert et al. (1995) will be employed, which is more effective than the conventional data driven approach selectors, such as cross-validation.

We consider the following model:

$$Y = \beta_1(U)X_1 + \beta_2(U)X_2 + \varepsilon, \tag{5.1}$$

$$P(\delta = 1|Y, U) = (1 + \exp(-v_0 - v_1(Y) - v_2(U)))^{-1}. \tag{5.2}$$

Under the model (5.1) and (5.2), we have that $E\{Y|U, X_1, X_2\} = \beta_1(U)X_1 + \beta_2(U)X_2$. A simple calculation yields

$$E\{Y|U, X_1, X_2, \delta = 1\} = \beta_1(U)X_1 + \beta_2(U)X_2 + \xi,$$

where

$$\xi = \frac{\sigma \int \tilde{y}(1 + \exp(-v_0 - v_1(\sigma\tilde{y} + \beta_1(U)X_1 + \beta_2(U)X_2) - v_2(U)))^{-1}\, d\Phi(\tilde{y})}{\int (1 + \exp(-v_0 - v_1(\sigma\tilde{y} + \beta_1(U)X_1 + \beta_2(U)X_2) - v_2(U)))^{-1}\, d\Phi(\tilde{y})}$$

and $\Phi(\cdot)$ is the standard normal cumulative distribution. Assume further $v_1(y) = \gamma y$. Therefore, it can be seen that $\xi \neq 0$ if $\gamma \neq 0$ because of the monotonicity of the exponential function. This means that unless $\gamma = 0$, a complete-data analysis leads to the considerable bias

$$E\{Y|U, X_1, X_2, \delta = 1\} - E\{Y|U, X_1, X_2\} = \xi \neq 0.$$

In this simulation, we choose $\beta_1(u) = 0.5\cos(5u)$, $\beta_2(u) = \exp(-(3u-1)^2) - 0.5$. The variables U, X_1 and X_2 follow a uniform distribution on [0,1], respectively. Furthermore, ε, U and (X_1, X_2) are independent and ε follows a normal distribution with mean zero and variance $\sigma^2 = 0.09$.

Firstly, we compare analyses based on $\hat{\beta}_{CC,k}(\cdot)$, $\hat{\beta}_{W,k}(\cdot, \pi)$, $\hat{\beta}_{WP,k}(\cdot, \hat{\pi})$, $\hat{\beta}_{WS,k}(\cdot, \hat{\pi})$. We consider the biases and variances of the different estimators above. We also choose the following two missing data mechanisms:

Case 1: $v_0 = -0.5, \quad v_1(Y) = 2Y, \quad v_2(U) = 1.5U.$

Case 2: $v_0 = 0.5, \quad v_1(Y) = Y - Y^2, \quad v_2(U) = 0.$

Approximately 45% of the data are missing under the above selection probabilities. For each cases, we conduct 2000 simulations with sample size 500. Figs. 1–5 summarize the results for the two cases. The true functions for $\beta_k(\cdot)$ $(k = 1, 2)$ are shown in Fig. 1 and the empirical biases of the estimators are shown in Figs. 2 and 4 for two cases. We can see from Figs. 2 and 4 that the complete-case data analysis has considerable biases. In Case 1, where the selection probabilities are correctly specified, $\hat{\beta}_{W,k}(\cdot, \pi)$, $\hat{\beta}_{WP,k}(\cdot, \hat{\pi})$ and $\hat{\beta}_{WS,k}(\cdot, \hat{\pi})$ have small biases. However, $\hat{\beta}_{WP,k}(\cdot, \hat{\pi})$ leads to larger biases than $\hat{\beta}_{W,k}(\cdot, \pi)$ and $\hat{\beta}_{WS,k}(\cdot, \hat{\pi})$ if the selection probabilities are incorrectly modelled as in Case 2. $\hat{\beta}_{WS,k}(\cdot, \hat{\pi})$ has a little larger biases than $\hat{\beta}_{W,k}(\cdot, \pi)$, which shows that estimating

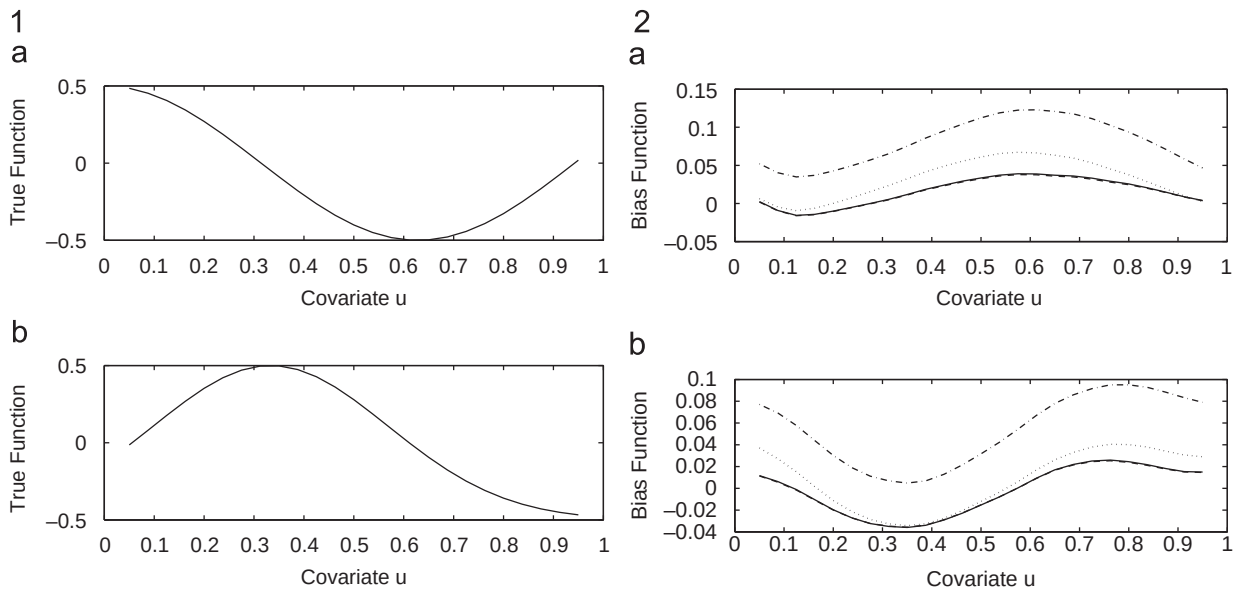

Fig. 1. True functions $\beta_k(u)$, $k=1,2$ in simulation study.

Fig. 2. Simulation study for biases from estimating $\beta_k(u)$, $k=1,2$, in Case 1. The solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dash-dotted curves are based on the estimate $\hat{\beta}_{CC,k}(u)$, the dotted curves are based on the estimate $\hat{\beta}_{WS,k}(u,\hat{\pi})$, and the dashed curves are based on the estimate $\hat{\beta}_{WP,k}(u,\hat{\pi})$, $k=1,2$.

selection probabilities nonparametrically has an effect on the bias in general. Figs. 3 and 5 show the sample variances of $\hat{\beta}_{W,k}(u,\pi)$, $\hat{\beta}_{WP,k}(u,\hat{\pi})$, and $\hat{\beta}_{WS,k}(u,\hat{\pi})$ and the relative efficiency curves of $\hat{\beta}_{WS,k}(u,\hat{\pi})$ to $\hat{\beta}_{W,k}(u,\pi)$ or $\hat{\beta}_{WP,k}(u,\hat{\pi})$. It is easily seen that $\hat{\beta}_{WS,k}(u,\hat{\pi})$ has a smaller variance than $\hat{\beta}_{W,k}(u,\pi)$ and $\hat{\beta}_{WP,k}(u,\hat{\pi})$ even if the selection probabilities are correctly specified, as discussed in Section 3.2.

A referee suggests to analyze the following case: the selection probabilities depends on whether or not Y exceeds a certain value, say zero. It is an interesting problem because that the assumption (f) may be violated. For this case, to further evaluate the performance of our proposed method, we consider the following missing mechanism:

$$Case\ 3: \pi(Y,U) = P\{\delta=1|Y,U\} = \Phi(I(Y>0)-0.5),$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution and $I(\cdot)$ is an indicator function. Under this missing mechanism, approximately 40% data are missing and the selection probabilities are 0.6915 or 0.3985, which implies that the continuous assumption on $\pi(\cdot)$ is violated. Figs. 6 and 7 summarize the results for this case. The empirical biases of the estimators are described in Fig. 6, which show that the complete-case data analysis has larger biases than our proposed method. Fig. 7 presents the sample variances of $\hat{\beta}_{W,k}(u,\pi)$, $\hat{\beta}_{WS,k}(u,\hat{\pi})$ and implies that $\hat{\beta}_{WS,k}(u,\hat{\pi})$ has a smaller variance than $\hat{\beta}_{W,k}(u,\pi)$, as expected. This kind of the selection probabilities depends only on the symbol of the response variable Y, that is, $\pi(y,u)$ is not continuous at the point $y=0$. Although the assumption (f) is not satisfied in this case, the performance is relatively well from Figs. 6 and 7. It is very surprising.

Finally, to demonstrate the power of the proposed bootstrap test, we consider the following simple model:

$$Y = \beta(U)X + \varepsilon,$$
$$P(\delta=1|Y,U) = (1+\exp(-Y))^{-1},$$

where U, X follow a uniform distribution on [0,1], respectively. Furthermore, ε, U and X are independent. ε follows a normal distribution with mean zero and variance $\sigma^2 = 0.25$. Approximately 45% of the data are missing under the above selection probabilities. Null hypothesis is as follows:

$$H_0: \beta(u) = 0.5, \quad u \in [0,1],$$

namely a linear model, against the alternative

$$H_a: \beta(u) = 0.5 + \alpha(u^2 - 0.5), \quad (0 \leqslant \alpha \leqslant 1).$$

We applied the goodness of fit in a simulation with 300 replications. For each replication, we generate 500 samples and repeat bootstrap sampling 300 times. Fig. 8 plots the simulated power function against α. When α=0, the specified alternative hypothesis

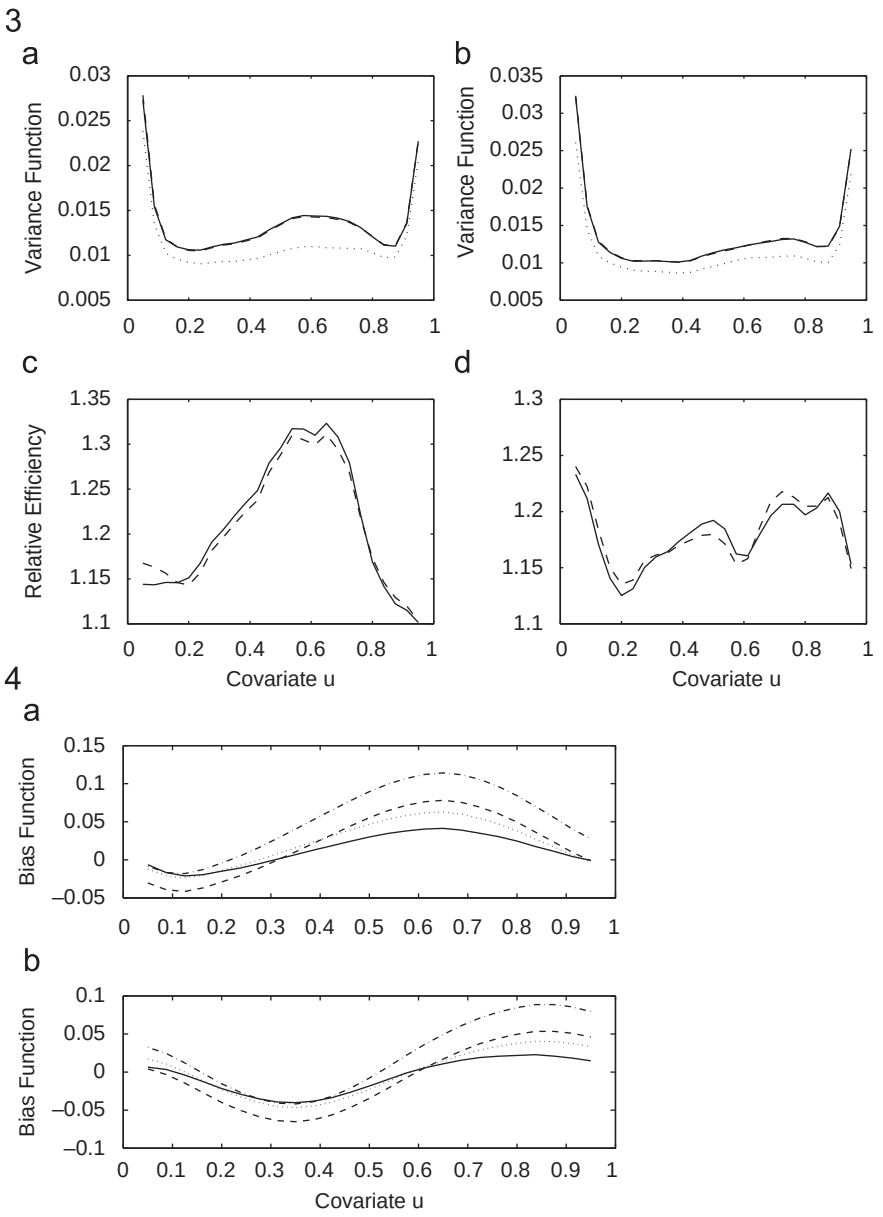

Fig. 3. Simulation study for variances and relative efficiency curves from estimating $\beta_k(u)$, $k=1,2$, in Case 1. For (a) and (b), the solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dotted curves are based on the estimate $\hat{\beta}_{WS,k}(u,\hat{\pi})$, and the dashed curves are based on the estimate $\hat{\beta}_{WP,k}(u,\hat{\pi})$, $k=1,2$. For (c) and (d), the solid curves are based on the ratio $var(\hat{\beta}_{W,k}(u,\pi))/var(\hat{\beta}_{WS,k}(u,\hat{\pi}))$, and the dashed curves are based on the ratio $var(\hat{\beta}_{WP,k}(u,\hat{\pi}))/var(\hat{\beta}_{WS,k}(u,\hat{\pi}))$.

Fig. 4. Simulation study for biases from estimating $\beta_k(u)$, $k=1,2$, in Case 2. The solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dash-dotted curves are based on the estimate $\hat{\beta}_{CC,k}(u)$, the dotted curves are based on the estimate $\hat{\beta}_{WS,k}(u,\hat{\pi})$, and the dashed curves are based on the estimate $\hat{\beta}_{WP,k}(u,\hat{\pi})$, $k=1,2$.

collapses into the null hypothesis. The size is 0.0433, which is close to the significant level of 5%. This demonstrate the bootstrap estimate of the null distribution is approximately correct. The power curve shows that our test is very powerful.

6. Data analysis of an AIDS clinic trial group study

In this section, we apply our method to the analysis of an AIDS clinical trial group (ACTG 315) study. The primary purpose of this study is to investigate the relationship between virologic and immunologic responses in AIDS clinical trials. It is well known that CD4+ cells are targets of HIV and decline to a lower level after HIV infection. Thus, when antiviral therapies suppress viral load, CD4+ cell counts may recover to a higher level (Lederman et al., 1998). In general, it is believed that the virologic response

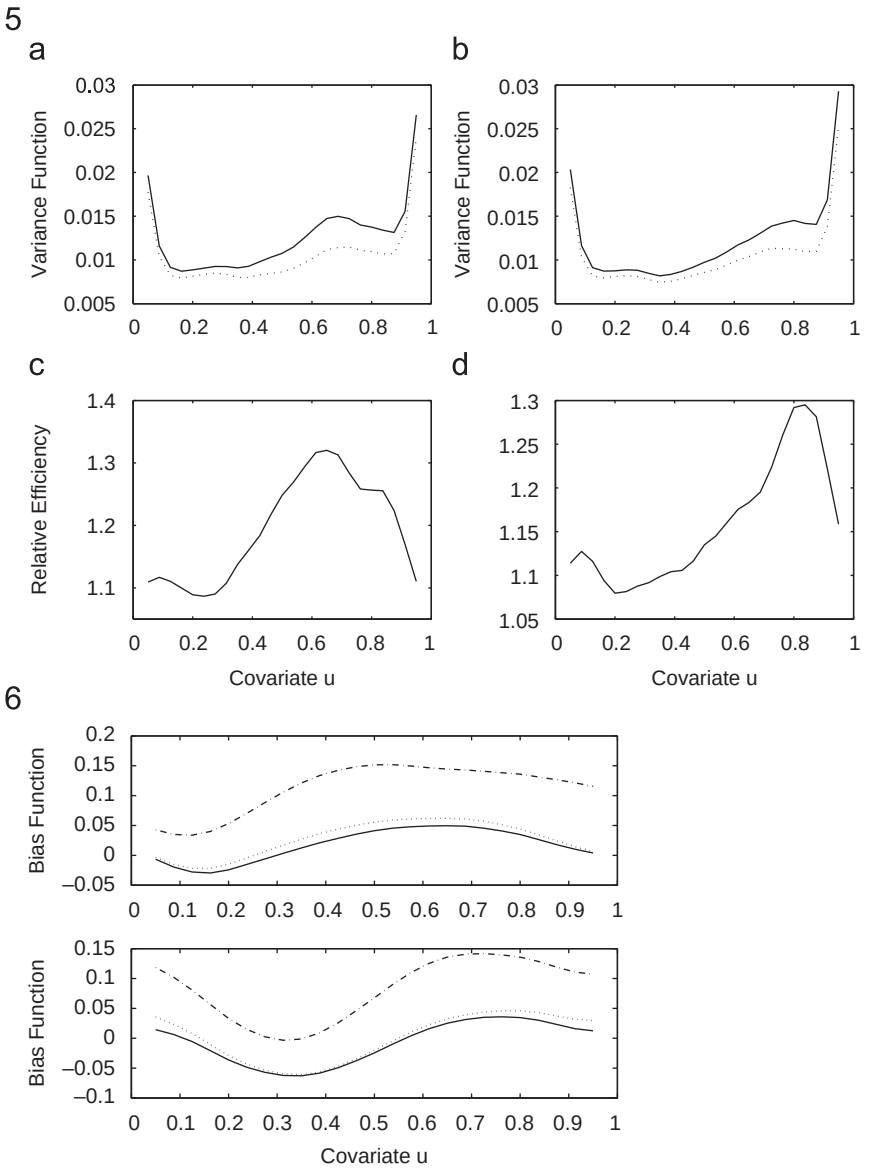

Fig. 5. Simulation study for variances and relative efficiency curves from estimating $\beta_k(u)$, k=1, 2, in Case 2. For (a) and (b), the solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dotted curves are based on the estimate $\hat{\beta}_{WS,k}(u,\hat{\pi})$, $k=1,2$. For (c) and (d), the solid curves are based on the ratio $var(\hat{\beta}_{W,k}(u,\pi))/var(\hat{\beta}_{WS,k}(u,\hat{\pi}))$.

Fig. 6. Simulation study for biases from estimating $\beta_k(u)$, $k=1,2$, in Case 3. The solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dash-dotted curves are based on the estimate $\hat{\beta}_{CC,k}(u)$, the dotted curves are based on the estimate $\hat{\beta}_{WS,k}(u,\hat{\pi})$, $k=1,2$.

(measured by viral load) and immunologic response (measured by CD4+ cell counts) are negatively correlated during antiviral treatments. However, their relationship may not be constant during the whole period of treatment. Liang et al. (2003) proposed the varying-coefficient models to study the relationship between virologic and immunologic responses.

In this study, both viral load and CD4+ cell counts were scheduled to be measured after initiation of an antiviral therapy. A total of 398 observations from 48 patients with 16.08% of CD4+ cell counts missing are obtained. Most of the missing values of the covariate CD4+ cell counts occurred because the covariate and the viral load were measured at different times. In other words, a miss in the data does not depend on the values being missing, and in this sense, is MAR (Little and Rubin, 1987). As Wu (2002) stated, the MAR assumption should be reasonable for this study.

To illustrate our method, we consider the following varying-coefficient model:

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})x_{ij} + \varepsilon_{ij}, \quad j = 1,\ldots,n_i, \quad i = 1,\ldots,n, \tag{6.1}$$

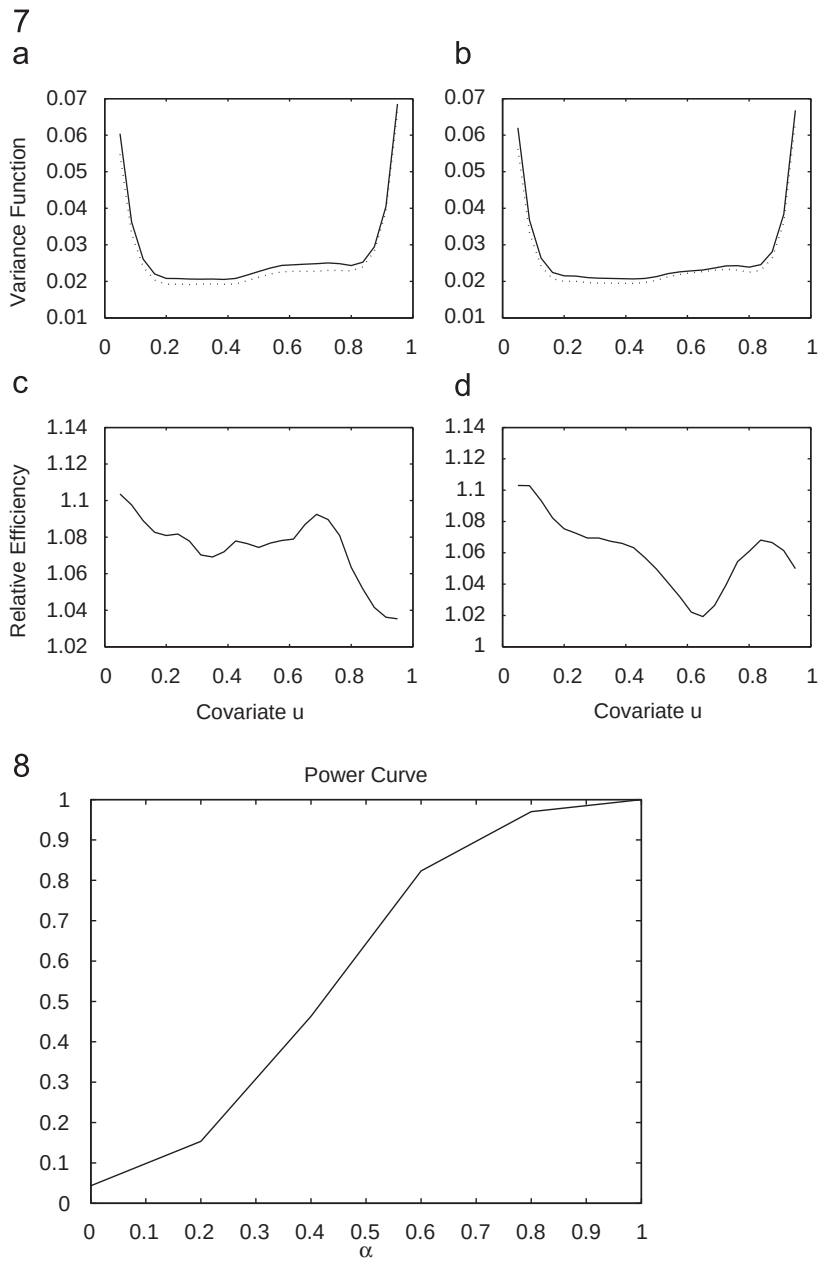

Fig. 7. Simulation study for variances and relative efficiency curves from estimating $\beta_k(u)$, k=1, 2, in Case 3. For (a) and (b), the solid curves are based on the estimate $\hat{\beta}_{W,k}(u,\pi)$, the dotted curves are based on the estimate $\tilde{\beta}_{WS,k}(u,\hat{\pi})$, $k=1,2$. For (c) and (d), the solid curves are based on the ratio $var(\hat{\beta}_{W,k}(u,\pi))/var(\hat{\beta}_{WS,k}(u,\hat{\pi}))$.
Fig. 8. The plot of power curve against α for the goodness-of-fit test.

where y_{ij} represents viral load, x_{ij} is CD4+ cell count at time t_{ij}. We assume $\{\varepsilon_{ij}, j=1,\ldots,n_i, i=1,\ldots,n\}$ are independent, that is, we ignore the dependence of $\{\varepsilon_{ij}, j=1,\ldots,n_i\}$ for each subject i.

We use the Epanechnikov kernel as in the simulation study and choose $h_0=0.6$ and $h=0.8$. For stable computation, we make the following transformation: $X=$(CD4+ cell counts)/100, and $t=$Day/30. We plot the estimates of $\beta_0(t)$ and $\beta_1(t)$ with their pointwise bootstrap standard error bands in Figs. 9 and 10. (We have tried different bandwidths h_0 and h, which showed similar results.) We can see that the estimates of $\beta_0(t)$ and $\beta_1(t)$ from different methods are very close, which is not too surprising, because only 16.08% of the CD4+ cell counts are missing. The results show that the viral load and CD4+ cell count responses are inversely related. The association between the virologic and immunologic responses, measured by $\beta_1(t)$, is stronger at the beginning of the treatment, but gradually dampens to become weak. However, after week 4, the association is recovered and becomes stronger and stronger.

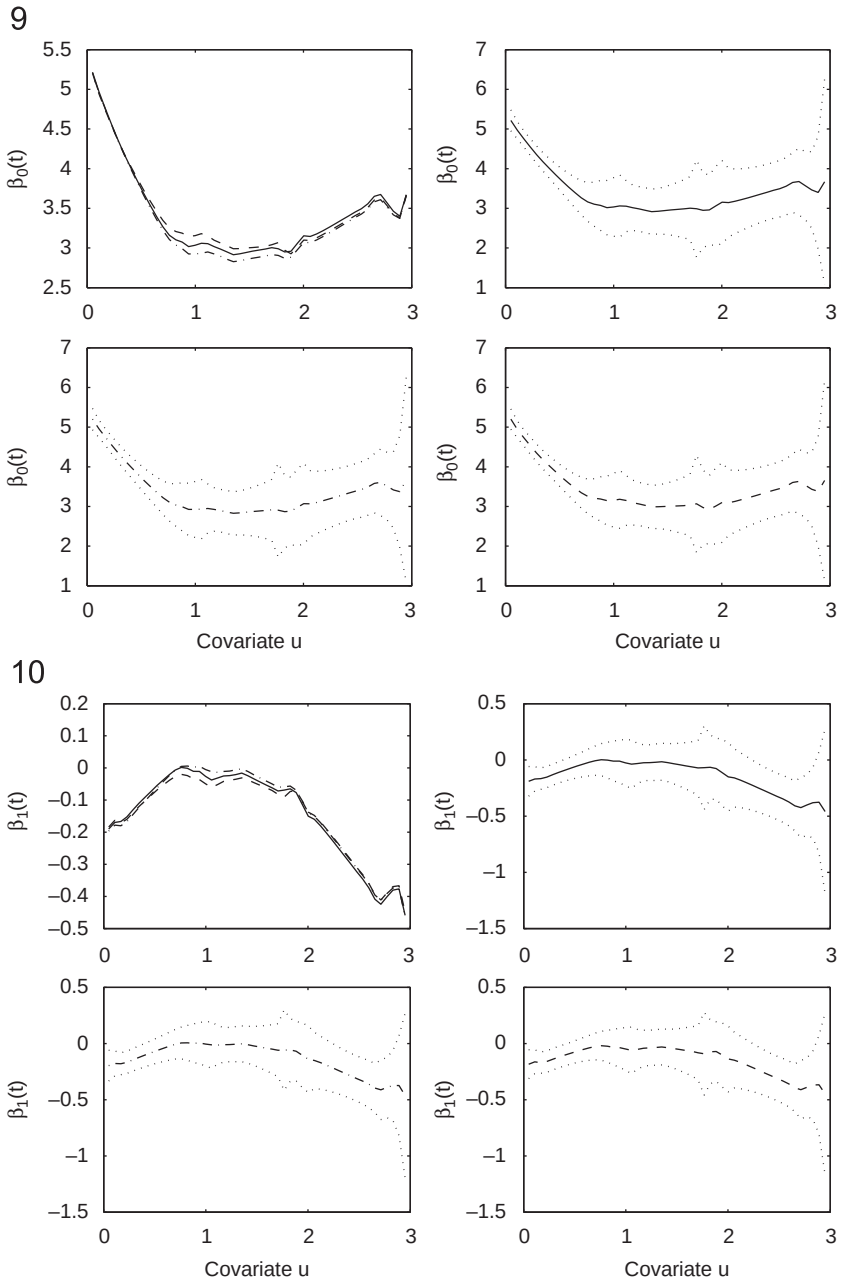

Fig. 9. Estimates and the corresponding pointwise confidence interval of $\beta_0(t)$ for the ACTG 315 dataset. The solid curves are based on the estimate $\hat{\beta}_{CC,0}(t)$, the dashed curves are based on the estimate $\hat{\beta}_{WP,0}(t,\hat{\pi})$, and the dash-dotted curves are based on the estimate $\hat{\beta}_{WS,0}(t,\hat{\pi})$.

Fig. 10. Estimates and the corresponding pointwise confidence interval of $\beta_1(t)$ for the ACTG 315 dataset. The solid curves are based on the estimate $\hat{\beta}_{CC,1}(t)$, the dashed curves are based on the estimate $\hat{\beta}_{WP,1}(t,\hat{\pi})$, and the dash-dotted curves are based on the estimate $\hat{\beta}_{WS,1}(t,\hat{\pi})$.

Finally, we apply the proposed bootstrap goodness-of-fit test to test whether the association between viral load and CD4+ cell responses is a function of treatment time t or a constant, that is, we test the hypothesis:

$$H_0: \beta_1(\cdot) = \text{Constant} \quad \text{versus} \quad H_a: \beta_1(\cdot) \neq \text{Constant}$$

under null hypothesis H_0, the model (6.1) becomes a partially linear model. The weighted residual sum of squares $WRSS_1$ for the model (6.1) is 0.4703 (0.4844) in contrast to $WRSS_0$ 0.4912 (0.4912) for the partially linear model under bandwidth $h_0 = 0.6, h = 0.6$ ($h_0 = 0.6, h = 0.8$). With 1000 bootstrap resampling, the p-value for this test is 0.0290 (0.0430), respectively, which rejects null hypothesis H_0. This suggests that the association between viral load and CD4+ cell responses is time varying.

7. Discussion

In this article, we suggest a locally weighted estimation approach with the selection probabilities estimated nonparametrically to fit the varying-coefficient models with missing covariates at random. Our theory demonstrates that locally weighted estimation approach with the selection probabilities estimated nonparametrically gives a smaller variance than that with the true selection probabilities. We also derive the estimators of coefficient functions based on simple locally augmented weighted estimating equation (SLAWEE). However, we show that the estimators based on SLAWEE do not improve the efficiency.

We suggest a wild bootstrap method to test if there exists a parametric structure for coefficient functions. This method is based on the comparison of the sum of the weighted residual squares under null and alternative models. Our simulation results show that this test method is very powerful. The methodology is applied to the ACTG(315) data.

A referee has mentioned that it would be an interesting problem to estimate the coefficient functions when U is really the vector of X_i's. It becomes some kinds of additive models or single-index varying-coefficient models. Our proposed method in this article may be also applied in this case, but the asymptotic properties of the estimated coefficient functions may be different from those studied in this article. Actually, the key point in the proofs is the quasi-linear structure for varying-coefficient models given the index variables U. We can make further research for this case.

Acknowledgement

The research of Heung Wong was supported by a grant from The Hong Kong Polytechnic University Research Committee. Min Chen's research work was supported by a grant from the Major State Basic Research Development Program of China (973 Program) (No. 2007CB814902), the National High Technology Research and Development Program of China (863 Program) (No. 2007AA12Z04), public-spirited Program of the Ministry of Water Resources of the People's Republic of China (No. 200801027) and the National Natural Science Foundation of China (No. 10628104, No.10721101).

Appendix

We first impose some assumptions on the regression model of the following theorems:

(a) $E(|\varepsilon|^4|U, X_1, \ldots, X_p) < \infty, E(|X_j|^{2s}) < \infty$, for some $s > 2, j = 1, \ldots, p$.
(b) $a_j''(\cdot)$ is continuous in a neighborhood of u_0, for $j = 1, \ldots, p$. Further, assume $a_j''(u_0) \neq 0$, for $j = 1, \ldots, p$.
(c) $r_{ij}''(\cdot)$ is continuous in a neighborhood of u_0, $r_j''(u_0) \neq 0$, for $j = 1, \ldots, p$, where $r_{ij}(u) = E(X_i X_j|U = u)$.
(d) The marginal density $f(u)$ of U has a continuous second derivative in some neighborhood of u_0 and $f(u_0) \neq 0$. And, $\sigma^2(u)$ is continuous at point u_0.
(e) The function $K(\cdot)$ is a bounded symmetric density function with a compact support.
(f) The probability function $\pi(y, u) > 0$ on the support of (Y, U) and has a bounded continuous second derivative.
(g) The bandwidth $h \to 0$, $nh \to \infty$ and $nh^\gamma/\ln(1/h) \to \infty$ for any $\gamma > s/(s-2)$ with s given in condition (a).
(h) $h_0/h \to 0$, $nh_0 \to \infty$.
(h′) The bandwidth $h \to 0$, $nh \to \infty$ and $nh^\gamma/\ln(1/h) \to \infty$ for any $\gamma > s/(s-2)$ with s given in condition (a). $\lambda_1 = O(\lambda_2)$, $\lambda_1 \to 0$, $n\lambda_1^2 h^2 \to \infty$ and $nh^{-1}\lambda_1^4 \to \infty$.

Remark 1. For example, $h = O(n^{-1/5})$, $h_0 = O(n^{-1/4})$, $\lambda_1 = \lambda_2 = O(n^{-1/6})$ satisfy the condition (h′).

The following notation will be used in the proofs of the theorems. Let

$$S = \Omega_p \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix}, \quad A = I_p \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product and I_p is $p \times p$ identity matrix. Denote $\tilde{S}$ be the matrix similar to S except replacing μ_i by v_i. Denote $E^x(Y)$ be condition expectation of Y given $X = x$ and $\hat{E}^x(Y)$ be local linear estimator of $E^x(Y)$.

Lemma 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d bivariate random vectors, where the Y_i's are scalar random variables. Assume further that $E|Y_1|^3 < \infty$ and $\sup_x \int |y|^s f(x, y)\, dy < \infty$, where $f(\cdot)$ denotes the joint density of (X, Y). Let $K(\cdot)$ be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_x \left| n^{-1} \sum_{i=1}^{n} \{K_h(X_i - x)Y_i - E[K_h(X_i - x)Y_i]\} \right| = O_p\left[\left\{ \frac{nh}{\ln(1/h)} \right\}^{-1/2} \right]$$

provided that $n^{2\varepsilon - 1}h \to \infty$ for some $\varepsilon < 1 - s^{-1}$.

Proof of Lemma 1. This follows immediately from the result obtained by Mack and Silverman (1982), see also Lemma 1 of Fan and Zhang (1999). □

Proof of Theorem 1. This follows easily from the result obtained by Theorem 3 of Fan and Zhang (1999). □

Proof of Theorem 2. Let $\Gamma_i = (1, Y_i, U_i)$, $D_n = n^{-1}\sum_{i=1}^n (\delta_i(1-\pi_i)/\pi_i)\tilde{X}_i\varepsilon_i$. Note $\hat{\alpha}$ is the maximum likelihood estimator of α. We have that $\hat{\alpha} - \alpha = O_p(n^{-1/2})$ and $\hat{\pi}_i - \pi_i = \pi_i(1-\pi_i)\Gamma_i(\hat{\alpha}-\alpha) + o_p(n^{-1/2})$.

Denote

$$\Xi = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i, \quad \Pi = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(U_i - u_0)\tilde{X}_i^\tau Y_i,$$

$$\tilde{\Xi} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i, \quad \tilde{\Pi} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau Y_i.$$

Then,

$$\begin{aligned}
\tilde{\Xi} &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i^2}(\hat{\pi}_i - \pi_i) K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i + o_p(n^{-1/2}) \\
&= \Xi - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i}(1-\pi_i) K_h(U_i - u_0)\tilde{X}_i^\tau \tilde{X}_i \Gamma_i(\hat{\alpha}-\alpha) + o_p(n^{-1/2}) \\
&= \Xi + O_p(n^{-1/2})
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\Pi} &= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau Y_i \\
&= \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i} K_h(U_i - u_0)\tilde{X}_i^\tau Y_i - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i^2}(\hat{\pi}_i - \pi_i) K_h(U_i - u_0)\tilde{X}_i^\tau Y_i + o_p(n^{-1/2}) \\
&= \Pi - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\pi_i}(1-\pi_i) K_h(U_i - u_0)\tilde{X}_i^\tau Y_i \Gamma_i(\hat{\alpha}-\alpha) + o_p(n^{-1/2}) \\
&= \Pi + O_p(n^{-1/2}).
\end{aligned}$$

Therefore, we can see that $\hat{\beta}_{WP,k}(u_0, \hat{\pi})$ has the same limiting distribution as $\hat{\beta}_{W,k}(u_0, \pi)$. □

Proof of Theorem 3. This proof is followed by the idea of Wang et al. (1997) and Fan and Zhang (1999).

Note that by Taylor's expansion, we have

$$Y_i = \tilde{X}_i^\tau \theta + \frac{1}{2}\sum_{j=1}^p [\beta_j''(\xi_{ij})(U_i - u_0)^2 X_{ij}] + \varepsilon_i,$$

where $\theta = (\beta_1(u_0), \beta_1'(u_0), \ldots, \beta_p(u_0), \beta_p'(u_0))^\tau$, $\tilde{X}_i = (X_{i1}, X_{i1}(U-u_0), \ldots, X_{ip}, X_{ip}(U-u_0))^\tau$, and ξ_{ij} is between U_i and u_0 for $j = 1, \ldots, p$. Thus,

$$\hat{\beta}_{WS,k}(u_0) = \beta_k(u_0) + \frac{1}{2}\sum_{j=1}^p e_{2k-1,2p}^\tau I_{1j} + e_{2k-1,2p}^\tau I_2,$$

where

$$I_{1j} = \tilde{\Xi}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau \beta_j''(\xi_{ij})(U_i - u_0)^2 X_{ij} \equiv \tilde{\Xi}^{-1}\tilde{I}_{1j}, \quad j = 1, \ldots, p,$$

$$I_2 = \tilde{\Xi}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^\tau \varepsilon_i \equiv \tilde{\Xi}^{-1}\tilde{I}_2$$

and

$$\tilde{\Xi} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^{\tau}\tilde{X}_i.$$

By calculating the mean and variance, one can easily get

$$\tilde{\Xi} = \Xi(1 + o_p(1)) = f(u_0)ASA(1 + o_p(1)).$$

Similarly, we have

$$\begin{aligned}\tilde{I}_{1j} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^{\tau}\beta_j''(\xi_{ij})(U_i - u_0)^2 X_{ij} \\ &= f(u_0)h^2\beta_j''(u_0)A\left(\alpha_j^{\tau}(u_0)\otimes(\mu_2, 0)\right)(1 + o_p(1)), \quad j = 1, \ldots, p.\end{aligned}$$

Now, we consider $\tilde{I}_2$. Denote $\hat{f}(w) = (nh_0^2)^{-1}\sum_{k=1}^{n} L_{h_0}(W_k - w)$.

$$\begin{aligned}\tilde{I}_2 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i^2}(\hat{\pi}_i - \pi_i)K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i + R_n \\ &= \tilde{I}_{21} - \frac{1}{n^2}\sum_{i=1}^{n}\delta_i\left(\frac{\sum_{j=1}^{n}(\delta_j - \pi_i)L_{h_0}(W_j - W_i)K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i}{h_0^2\pi_i^2\hat{f}(W_i)}\right) + R_n \\ &= \tilde{I}_{21} - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\frac{(\delta_j - \pi_i)(\delta_i - \pi_i)L_{h_0}(W_j - W_i)K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i}{h_0^2\pi_i^2\hat{f}(W_i)}\right) \\ &\quad - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\frac{\pi_i(\delta_j - \pi_i)L_{h_0}(W_j - W_i)K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i}{h_0^2\pi_i^2\hat{f}(W_i)}\right) + R_n \\ &= \tilde{I}_{21} - \tilde{I}_{22} - \tilde{I}_{23} + R_n,\end{aligned}$$

where

$$R_n = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\pi_i^3}(\hat{\pi}_i - \pi_i)^2 K_h(U_i - u_0)\tilde{X}_i^{\tau}\varepsilon_i(1 + o_p(1)),$$

with mean $O(h_0^4)$ and variance $o((nh)^{-1})$. It is easily seen that

$$\tilde{I}_{23} = \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j - \pi_j}{\pi_j} K_h(U_j - u_0)E\{\tilde{X}^{\tau}\varepsilon|W_i\} + R_{1n},$$

where R_{1n} with mean $O(h_0^2)$ and variance $o((nh)^{-1})$. Next, we can prove that $\tilde{I}_{22}$ has mean $O(h_0^2)$ and variance $o((nh)^{-1})$ similar to Wang et al. (1997). Thus,

$$\tilde{I}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi_i}\tilde{X}_i^{\tau}\varepsilon_i - \frac{\delta_i - \pi_i}{\pi_i}E\{\tilde{X}^{\tau}\varepsilon|W_i\}\right)K_h(U_i - u_0) + \tilde{R}_n,$$

where $\tilde{R}_n$ with mean $O(h_0^2)$ and variance $o((nh)^{-1})$. In particular,

$$var(\tilde{I}_2) = \frac{1}{nh}\tilde{\Omega}_p^{\star}\otimes I_2(1 + o(1)).$$

By Lemma 1 and the condition $h_0/h \to 0$, we obtain that the asymptotic bias of $\hat{\beta}_{WS,k}(u_0, \hat{\pi})$ is given by

$$\begin{aligned}bias(\hat{\beta}_{WS,k}(u_0, \hat{\pi})) &= \frac{1}{2}h^2\sum_{j=1}^{p}\beta_j''(u_0)e_{2k-1,2p}^{\tau}S^{-1}(\alpha_j^{\tau}(u_0)\otimes(\mu_2, 0))(1 + o(1)) \\ &= \frac{1}{2}h^2\mu_2\beta_k''(u_0)(1 + o(1))\end{aligned}$$

and the asymptotic variance of $\hat{\beta}_{WS,k}(u_0,\hat{\pi})$ is given by

$$var(\hat{\beta}_{WS,k}(u_0,\hat{\pi})) = \frac{v_0}{nhf(u_0)} e_{k,p}^{\tau}\Omega_p^{-1}\tilde{\Omega}_p^{\star}\Omega_p^{-1}e_{k,p}(1+o_p(1)).$$

It can be shown by checking Lyapunov's condition that $\hat{\beta}_{WS,k}(u_0,\hat{\pi})$ is asymptotically normally distributed. The proof is completed. □

Before proving Theorem 4, we present one lemma which will be used in the proofs.

Lemma 2. *Under the assumptions* (a)–(f) *and* (h′),

$$E(C_n)=o(h^2),\quad var(C_n)=o((nh)^{-1}),$$

$$E(G_{n,j})=o(h^2),\quad var(G_{n,j})=o((nh)^{-1})$$

and

$$E(H_{n,j})=o(h^2),\quad var(H_{n,j})=o((nh)^{-1}),\ j=1,\dots,p,$$

where

$$C_n=\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}[\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau})Y_i-\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}Y)],$$

$$G_{n,j}=\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}[\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}a_j''(\xi_j)(U-u_0)^2X_j)-E^{Y_i,U_i}(\tilde{X}^{\tau}a_j''(\xi_j)(U-u_0)^2X_j)],$$

$$H_{n,j}=\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}[\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}\varepsilon)-E^{Y_i,U_i}(\tilde{X}^{\tau}\varepsilon)],\quad j=1,\dots,p.$$

Proof of Lemma 2. This proof is followed by the idea of Wang et al. (1998). Here, we only prove $E(C_n)=o(h^2)$ and $var(C_n)=o((nh)^{-1})$ due to the similarity of $G_{n,j}$ and $H_{n,j}$.

Denote $\tilde{E}_i=\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau})Y_i$, $\hat{E}_i=\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}Y)$, and $E_i=E^{Y_i,U_i}(\tilde{X}^{\tau}Y)$. First, note that $E\{(\hat{E}_i-E_i)|Y_i,U_i\}=\lambda_1^2c_1(Y_i,U_i)\{1+o_p(1)\}$, and $var\{(\hat{E}_i-E_i)|Y_i,U_i\}=(n\lambda_1^2)^{-1}c_2(Y_i,U_i)\{1+o_p(1)\}$, $E\{(\tilde{E}_i-E_i)|Y_i,U_i\}=\lambda_1^2c_3(Y_i,U_i)\{1+o_p(1)\}$, and $var\{(\tilde{E}_i-E_i)|Y_i,U_i\}=(n\lambda_1^2)^{-1}c_4(Y_i,U_i)\{1+o_p(1)\}$, for some function c_1, c_2, c_3 and c_4.

Denote $C_{n1}=(1/n)\sum_{i=1}^{n}K_h(U_i-u_0)((\delta_i-\pi_i)/\pi_i)[\hat{E}_i-E_i]$, $C_{n2}=(1/n)\sum_{i=1}^{n}K_h(U_i-u_0)((\delta_i-\pi_i)/\pi_i)[\tilde{E}_i-E_i]$. Let $\hat{E}_{i(j)}$ denote $\hat{E}_i$ without using subject j. Then

$$\begin{aligned}E(C_{n,1})&=E\left[K_h(U_1-u_0)\frac{\delta_1-\pi_1}{\pi_1}(\hat{E}_1-E_1)\right]\\&=E\left\{E\left[K_h(U_1-u_0)\frac{\delta_1-\pi_1}{\pi_1}(\hat{E}_{1(1)}-E_1)|Y_1,U_1\right]\right\}+O((n\lambda_1^2)^{-1})\\&=E\left\{K_h(U_1-u_0)E\left[\frac{\delta_1-\pi_1}{\pi_1}|Y_1,U_1\right]E[(\hat{E}_{1(1)}-E_1)|Y_1,U_1]\right\}+O((n\lambda_1^2)^{-1})\\&=O((n\lambda_1^2)^{-1})=o(h^2).\end{aligned}$$

Denote $S_i=K_h(U_i-u_0)((\delta_i-\pi_i)/\pi_i)(\hat{E}_i-E_i)$. Similar to the above, we can easily obtain that $cov(S_i,S_k)=O((n^2\lambda_1^4)^{-1})$ when $i\neq k$. Then the variance of the k-th element of $G_{n,j}$ is

$$\begin{aligned}var[(C_{n,1})_k]&=\frac{1}{n}var\left[K_h(U_1-u_0)\frac{\delta_1-\pi_1}{\pi_1}(\hat{E}_1-E_1)_k\right]+O(n^2\lambda_1^4)^{-1}\\&=\frac{1}{n}E\left[K_h^2(U_1-u_0)\frac{1-\pi_1}{\pi_1}var\{(\hat{E}_{1(1)}-E_1)_k|Y_1,U_1\}\right]\\&\quad+\frac{1}{n}var\left[K_h(U_1-u_0)\frac{\delta_1-\pi_1}{\pi_1}E\{(\hat{E}_{1(1)}-E_1)_k|Y_1,U_1\}\right]+O(n^2\lambda_1^4)^{-1}\\&=O((nh)^{-1}(n\lambda_1^2)^{-1}+(nh)^{-1}\lambda_1^4)+O(n^2\lambda_1^4)^{-1}\\&=o((nh)^{-1}).\end{aligned}$$

Therefore, we obtain that $var(C_{n1})=o((nh)^{-1})$. Similarly, we can get that $E(C_{n2})=o(h^2)$, $var(C_{n2})=o((nh)^{-1})$. Therefore, $E(C_n)=o(h^2)$, $var(C_n)=o((nh)^{-1})$. □

Proof of Theorem 4. Note that by Taylor's expansion, we have

$$Y=\tilde{X}^\tau\theta+\frac{1}{2}\sum_{j=1}^{p}[\beta_j''(\xi_j)(U-u_0)^2X_j]+\varepsilon,$$

where $\theta=(\beta_1(u_0),\beta_1'(u_0),\ldots,\beta_p(u_0),\beta_p'(u_0))^\tau$, $\tilde{X}=(X_1,X_1(U-u_0),\ldots,X_p,X_p(U-u_0))^\tau$ and ξ_j is between U and u_0 for $j=1,\ldots,p$. Thus,

$$\hat{\beta}_{A,k}(u_0,\pi)=\beta_k(u_0)+e_{2k-1,2p}^\tau I_1+\frac{1}{2}\sum_{j=1}^{p}e_{2k-1,2p}^\tau[I_{2j}-I_{3j}]+e_{2k-1,2p}^\tau[I_4-I_5],$$

where

$$I_1=\Xi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}(\hat{E}^{Y_i,U_i}(\tilde{X}^\tau)Y_i-\hat{E}^{Y_i,U_i}(\tilde{X}^\tau Y))\right]=\Xi^{-1}\tilde{I}_1,$$

$$I_{2j}=\Xi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i}{\pi_i}\tilde{X}_i^\tau\beta_j''(\xi_{ij})(U_i-u_0)^2X_{ij}\right]=\Xi^{-1}\tilde{I}_{2j},$$

$$I_{3j}=\Xi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^\tau\beta_j''(\xi_j)(U-u_0)^2X_j)\right]=\Xi^{-1}\tilde{I}_{3j},\quad j=1,\ldots,p,$$

$$I_4=\Xi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i}{\pi_i}\tilde{X}_i^\tau\varepsilon_i\right]=\Xi^{-1}\tilde{I}_4,$$

$$I_5=\Xi^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^\tau\varepsilon)\right]=\Xi^{-1}\tilde{I}_5,$$

$$\Xi=\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\left[\frac{\delta_i}{\pi_i}\tilde{X}_i^\tau\tilde{X}_i-\frac{\delta_i-\pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^\tau\tilde{X})\right].$$

(1) Assume π is known.

First, let us calculate the asymptotic bias of $\hat{\beta}_{A,k}(u_0,\pi)$. By calculating the mean and variance, one can get

$$\Xi=f(u_0)ASA(1+o_p(1)),$$

$$\tilde{I}_{2j}=f(u_0)h^2\beta_j''(u_0)A(\alpha_j\otimes(\mu_2,0)^\tau)(1+o_p(1)),\quad j=1,\ldots,p,$$

$$\tilde{I}_4=O_p((nh)^{-1/2}).$$

Next we calculate $\tilde{I}_1$, $\tilde{I}_{3j}$ and $\tilde{I}_5$. Through Lemmas 1 and 2, we obtain that

$$E(\tilde{I}_{3j})=E\left(\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}[\hat{E}_i-E_i]\right)+E\left(\frac{1}{n}\sum_{i=1}^{n}K_h(U_i-u_0)\frac{\delta_i-\pi_i}{\pi_i}E_i\right)$$
$$=o(h^2),$$

$$E(\tilde{I}_1)=o(h^2),\quad E(\tilde{I}_5)=o(h^2).$$

Therefore, the asymptotic bias of $\hat{\beta}_{A,k}(u_0,\pi)$ is given by

$$bias(\hat{\beta}_{A,k}(u_0,\pi))=\tfrac{1}{2}h^2\mu_2\beta_k''(u_0)(1+o(1)).$$

Using an asymptotic argument similar to the above, it is easy to calculate that the asymptotic variance of $\hat{\beta}_{A,k}(u_0,\pi)$ is given by

$$var(\hat{\beta}_{A,k}(u_0,\pi)) = \frac{v_0}{nhf(u_0)} e_{2k-1,2p}^{\tau} \Omega_p^{-1} \tilde{\Omega}_p^{\star} \Omega_p^{-1} e_{2k-1,2p}(1+o(1)),$$

where

$$\tilde{\Omega}_p^{\star} = E\left\{\frac{\varepsilon^2}{\pi} X^{\tau} X | U = u_0\right\} - E\left\{\frac{1-\pi}{\pi}[E(X^{\tau}\varepsilon|Y,U)]^{\otimes 2} | U = u_0\right\}.$$

(2) Assume that π is unknown and estimated by kernel smoothing as that of (2.12). Denote

$$\hat{\Xi} = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\left[\frac{\delta_i}{\hat{\pi}_i}\tilde{X}_i^{\tau}\tilde{X}_i - \frac{\delta_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}\tilde{X})\right],$$

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau})Y_i.$$

By Theorem 3, it is sufficient to prove that

$$\hat{\Xi} = \Xi(1+o_p(1)) = f(u_0)ASA(1+o_p(1)),$$

$$E(\hat{I}) = o(h^2), \quad var(\hat{I}) = o((nh)^{-1}).$$

It is easily seen that

$$\hat{\Xi} = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\left[\frac{\delta_i}{\pi_i}\tilde{X}_i^{\tau}\tilde{X}_i - \frac{\delta_i - \pi_i}{\pi_i}\hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}\tilde{X})\right] - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i(\hat{\pi}_i - \pi_i)}{\pi_i^2}[\tilde{X}_i^{\tau}\tilde{X}_i - \hat{E}^{Y_i,U_i}(\tilde{X}^{\tau}\tilde{X})](1+o_p(1))$$

$$= \Xi(1+o_p(1)) = f(u_0)ASA(1+o_p(1)).$$

Let $\hat{E}_i = \hat{E}^{Y_i,U_i}(\tilde{X}^{\tau})Y_i$, $E_i = E^{Y_i,U_i}(\tilde{X}^{\tau}Y)$. Then,

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i - \hat{\pi}_i}{\hat{\pi}_i}\hat{E}_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i - \pi_i}{\pi_i}(\hat{E}_i - E_i) + \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i - \pi_i}{\pi_i}E_i$$

$$-\frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_0)\frac{\delta_i(\hat{\pi}_i-\pi_i)}{\pi_i^2}(\hat{E}_i-E_i)(1+o_p(1)) - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_0)\frac{\delta_i(\hat{\pi}_i-\pi_i)}{\pi_i^2}E_i - \frac{1}{n}\sum_{i=1}^{n} K_h(U_i-u_0)\frac{\delta_i(\hat{\pi}_i-\pi_i)^2}{\pi_i^3}E_i(1+o_p(1))$$

$$= \hat{I}_1 + \hat{I}_2 - \hat{I}_3 - \hat{I}_4 - \hat{I}_5.$$

By Lemma 2, we obtain that

$$E(\hat{I}_1) = o(h^2), \quad var(\hat{I}_1) = o((nh)^{-1}).$$

Similar to $\tilde{I}_2$ of Theorem 3, we have

$$\hat{I}_4 = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\frac{\delta_i - \pi_i}{\pi_i}E_i + R_n,$$

where

$$E(R_n) = O(h_0^2), \quad var(R_n) = o((nh)^{-1}).$$

It is easily seen that

$$E(\hat{I}_3) = o(h^2), \quad var(\hat{I}_3) = o((nh)^{-1}),$$

$$E(\hat{I}_5) = O(h_0^4), \quad var(\hat{I}_5) = o((nh)^{-1}).$$

Then, we can get that

$$E(\hat{I}) = o(h^2), \quad var(\hat{I}) = o((nh)^{-1}).$$

Therefore, $\hat{\beta}_{A,k}(u_0,\hat{\pi})$ has the same asymptotic normal distribution as $\hat{\beta}_{A,k}(u_0,\pi)$. □

References

Cai, Z., Fan, J., Yao, Q., 2000. Functional-coefficient regression models for nonlinear time series. J. Amer. Statist. Assoc. 95, 941–956.
Carroll, R.J., Fan, J., Gijbels, I., Wand, M.P., 1997. Generalized partially linear single index models. J. Amer. Statist. Assoc. 92, 477–489.
Carroll, R.J., Ruppert, D., Welsh, A.H., 1998. Local estimating equation. J. Amer. Statist. Assoc. 93, 214–227.
Chiang, C.T., Rice, J.A., Wu, C.O., 2001. Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. J. Amer. Statist. Assoc. 96, 605–619.
Cleveland, W.S., Grosse, E., Shyu, W.M., 1991. Local regression models. In: Chambers, J.M., Hastie, T.J. (Eds.), Statistical Models in S. Wadsworth, Brooks–Cole, Pacific Grove, CA, pp. 309–376.
Copas, J.B., 1983. Plotting *p* against *x*. Appl. Statist. 32, 25–31.
Eubank, R.L., 1988. Smoothing Spline and Nonparametric regression. Marcel Dekker, New York.
Eubank, R.L., Huang, C., Maldonado, Y.M., Wang, N., Wang, S., Buchanan, R.J., 2004. Smoothing spline estimation in varying coefficient models. J. Roy. Statist. B 66, 653–667.
Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.
Fan, J., Härdle, W., Mammen, E., 1998. Direct estimation of additive and linear components for high dimensional data. Ann. Statist. 26, 943–971.
Fan, J., Zhang, J.T., 1999. Statistical estimation in varying coefficient models. Ann. Statist. 27, 1491–1518.
Fan, J., Zhang, J.T., 2000a. Two step estimation of functional linear models with applications to longitudinal data. J. Roy. Statist. Soc. Ser. B 57, 371–394.
Fan, J., Zhang, W.Y., 2000b. Simultaneous confidence bands and hypothesis testing in varying coefficient models. Scand. J. Statist. 27, 715–731.
Fan, J., Zhang, C.M., Zhang, J., 2001. Generalized likelihood ratio statistics and Wilks phenomenon. Ann. Statist. 29, 153–193.
Flander, W.D., Greenland, S., 1991. Analytic methods for two stage case–control studies and other stratified designs. Statist. Med. 10, 739–747.
Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). Ann. Statist. 19, 1–141.
Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London.
Gu, C., Wahba, G., 1993. Smoothing spline ANOVA with component-wise Bayesian 'confidence intervals'. J. Comput. Graph. Statist. 2, 97–117.
Härdle, W., Stoker, T.M., 1989. Investigating smooth multiple regression by the method of average derivatives. J. Amer. Statist. Assoc. 84, 986–995.
Hastie, T.J., Tibishirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.
Hastie, T.J., Tibishirani, R.J., 1993. Varying-coefficient models. J. Roy. Statist. Soc. Ser. B 55, 757–796.
Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.
Hoover, D.R., Rice, J.A., Wu, C.O., Yang, Y., 1998. Nonparametric smoothing estimates of time varying coefficient models with longitudinal data. Biometrika 85, 809–822.
Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663–685.
Huang, J., Wu, C.O., Zhou, L., 2002. Varying coefficient models and basis function approximations for the analysis of repeated measurements. Biometrika 89, 111–128.
Huang, J., Wu, C.O., Zhou, L., 2004. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. Statist. Sinica 14, 763–788.
le Cessie, S., van Houwelingen, F.C., 1991. A goodness-of-fit test for binary regression models, based on smoothing methods. Biometrics 47, 1267–1282.
Lederman, M.M., Connick, E., Landay, A., et al., 1998. Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine and ritonavir: results of AIDS clinical trials group protocol 315. J. Infect. Diseases 178, 70–79.
Li, K.C., 1991. Sliced inverse regression for dimension reduction (with discussion). J. Amer. Statist. Assoc. 86, 316–342.
Liang, H., Wang, S., Robins, J.M., Carroll, R.J., 2004. Estimation in partially linear models with missing covariates. J. Amer. Statist. Assoc. 99, 357–367.
Liang, H., Wu, H., Carroll, R.J., 2003. The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effects varying coefficient models with measurement error. Biostatistics 3, 297–312.
Little, R.J., 1992. Regression with missing *X*'s: a review. J. Amer. Statist. Assoc. 87, 1127–1137.
Little, R.J., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.
Mack, Y., Silverman, B., 1982. Weak and strong uniform consistency of kernel regression estimates. Z. Wahrsch. Verw. Gebiete 61, 405–415.
Rubin, D.B., 1976. Inference and missing data. Biomoetrika 63, 581–592.
Robins, J.M., Hsieh, F., Newey, W., 1995. Semiparametric estimation of a conditional density with missing or mismeasured covariates. J. Roy. Statist. Soc. Ser. B 57, 409–424.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89, 846–866.
Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. J. Amer. Statist. Assoc. 90, 1257–1270.
Tsiatis, A.A., 2006. Semiparametric Theory and Missing Data. Springer, New York.
Van der Laan, M.J., Rubins, J.M., 2003. Unified Methods for Censored Longitudinal Data and Causality. Springer, New York.
Wahba, G.,1984. Partial spline models for semiparametric estimation of functions of several variables. In Statistical Analysis of Time Series. Proceeding of the Japan-U.S. Joint Seminar, Tokyo 319–329. Institute of Statistical Mathematics, Tokyo.
Wang, C.Y., Wang, S., Gutierrez, R.G., Carroll, R.J., 1998. Local linear regression for generalized linear models with missing data. Ann. Statist. 26, 1028–1050.
Wang, C.Y., Wang, S., Zhao, L.P., Ou, S.T., 1997. Weighted semiparametric estimation in regression analysis with missing covariate data. J. Amer. Statist. Assoc. 92, 512–525.
Wu, C.O., Chiang, C.T., Hoover, D.R., 1998. Asymptotic confidence regions for kernel smoothing of a varying coefficient model with longitudinal data. J. Amer. Statist. Assoc. 93, 1388–1402.
Wu, C.O., Yu, K.F., Chiang, C.T., 2000. A two-step smoothing method for varying coefficient models with repeated measurements. Ann. Inst. Statist. Math. 52, 519–543.
Wu, L., 2002. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies. J. Amer. Statist. Assoc. 97, 955–964.
Zhang, W., Lee, S., Song, X., 2000. Local polynomial fitting in semivarying coefficient models. J. Multivarite Anal. 82, 166–188.
Zhao, L.P., Lipsitz, S., 1992. Design and analysis of two-stage studies. Statist. Med. 11, 769–782.