Biometrika (1998), **85**, 4, *pp.* 809–822
Printed in Great Britain

Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data

BY DONALD R. HOOVER

Department of Epidemiology, The Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

dhoover@share.sph.jhu.edu

JOHN A. RICE

Department of Statistics, University of California, Berkeley, California 94720, U.S.A.

rice@stat.berkeley.edu

COLIN O. WU

Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218, U.S.A.

colinwu@columbo.mts.jhu.edu

AND LI-PING YANG

Department of Epidemiology, The Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

lpyang@share.sph.jhu.edu

SUMMARY

This paper considers nonparametric estimation in a varying coefficient model with repeated measurements (Y_{ij}, X_{ij}, t_{ij}), for $i=1,\ldots,n$ and $j=1,\ldots,n_i$, where $X_{ij}=(X_{ij0},\ldots,X_{ijk})^{\mathrm{T}}$ and (Y_{ij}, X_{ij}, t_{ij}) denote the jth outcome, covariate and time design points, respectively, of the ith subject. The model considered here is $Y_{ij}=X_{ij}^{\mathrm{T}}\beta(t_{ij})+\varepsilon_i(t_{ij})$, where $\beta(t)=(\beta_0(t),\ldots,\beta_k(t))^{\mathrm{T}}$, for $k\geqslant 0$, are smooth nonparametric functions of interest and $\varepsilon_i(t)$ is a zero-mean stochastic process. The measurements are assumed to be independent for different subjects but can be correlated at different time points within each subject. Two nonparametric estimators of $\beta(t)$, namely a smoothing spline and a locally weighted polynomial, are derived for such repeatedly measured data. A crossvalidation criterion is proposed for the selection of the corresponding smoothing parameters. Asymptotic properties, such as consistency, rates of convergence and asymptotic mean squared errors, are established for kernel estimators, a special case of the local polynomials. These asymptotic results give useful insights into the reliability of our general estimation methods. An example of predicting the growth of children born to HIV infected mothers based on gender, HIV status and maternal vitamin A levels shows that this model and the corresponding nonparametric estimators are useful in epidemiological studies.

Some key words: Longitudinal data; Mean squared error; Nonparametric estimation; Rates of convergence; Varying coefficient model.

1. INTRODUCTION

In a longitudinal study, let $Y(t)$ and $X(t)$ be the real-valued outcome of interest and the R^{k+1}-valued column covariate vector, respectively, observed at time t, where $k \geqslant 0$. Suppose there are n subjects, and for the ith subject there are $n_i \geqslant 1$ repeated measurements of $(Y(t), X(t), t)$ over time. The jth observation of $(Y(t), X(t), t)$ for the ith subject is denoted by (Y_{ij}, X_{ij}, t_{ij}), for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$, where $X_{ij} \in R^{k+1}$ is given by the column vector $X_{ij} = (X_{ij0}, \ldots, X_{ijk})^{\mathrm{T}}$.

Under the classical linear model framework, theory and methods of regression with repeated observations have been extensively studied in the literature. These results include Pantula & Pollock (1985), Ware (1985), Diggle (1988), Jones & Ackerson (1990) and Jones & Boadi-Boteng (1991). A summary of different types of parametric approach can be found in Diggle, Liang & Zeger (1994, Ch. 4). While parametric approaches are useful, questions will always arise about the adequacy of the model assumptions and the potential impact of model misspecifications on the analysis. This motivates the use of nonparametric approaches.

For nonparametric models with fixed design time points, Hart & Wehrly (1986), Altman (1990) and Hart (1991) considered kernel methods for estimating the expectation, $E\{Y(t)\}$, without the presence of the covariate $X(t)$, and derived a class of generalised crossvalidation bandwidth selection procedures. Rice & Silverman (1991) considered a class of smoothing splines and proposed a method of choosing the smoothing parameters by crossvalidation in which subjects were left out one at a time. Although the existing kernel and spline methods are successful in predicting the mean curve of $Y(t)$ over time, they only consider the effect of t and do not take account of other possibly important covariates.

To quantify the influence of covariates, Zeger & Diggle (1994) and Moyeed & Diggle (1994) studied a semiparametric model

$$Y_{ij} = \mu(t_{ij}) + X_{ij}^{\mathrm{T}}\beta + \varepsilon_i(t_{ij}), \qquad (1{\cdot}1)$$

where $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is a vector of unknown constants in R^k, $\mu(t)$ is an arbitrary smooth function of t, and the error term $\varepsilon_i(t)$ is a zero-mean stochastic process. By generalising the methods of Hastie & Tibshirani (1990, Ch. 4), Zeger & Diggle (1994) and Moyeed & Diggle (1994) suggested a backfitting procedure which initially estimates $\mu(t)$ by a class of kernel estimators and then iteratively estimates β and $\mu(t)$.

We consider in this paper a direct generalisation of model (1·1) that allows the coefficients to vary over time:

$$Y_{ij} = X_{ij}^{\mathrm{T}}\beta(t_{ij}) + \varepsilon_i(t_{ij}), \qquad (1{\cdot}2)$$

where, for all $t \in R$, $\beta(t) = (\beta_0(t), \ldots, \beta_k(t))^{\mathrm{T}}$ $(k \geqslant 0)$ are arbitrary smooth functions of t, $\varepsilon_i(t)$ is a realisation of a zero-mean stochastic process $\varepsilon(t)$, and X_{ij} and ε_i are independent. We note that the process $\varepsilon_i(t)$ need not have zero mean for each subject. When the data are obtained from independent cross-sectional samples, (1·2) reduces to the varying coefficient models studied by Hastie & Tibshirani (1993). The linear procedures we develop in this paper do not involve attempting to model the correlation structure of $\varepsilon(t)$. We believe that in many applications accurate modelling of the correlation of $\varepsilon(t)$ is impracticable. Furthermore, it is quite conceivable that such an attempt could actually increase the mean squared error.

In § 2, we present the two computationally straightforward nonparametric estimators of $\beta(t)$, namely smoothing splines and locally weighted polynomials, and set forth crossvalidation criteria for selecting smoothing parameters. Section 3 applies our procedures to the prediction of growth of children born to HIV infected mothers, based on maternal vitamin A levels and children's gender and HIV status. Section 4 gives the asymptotic properties of the kernel estimators. In § 5 we summarise our results and discuss some related issues involving estimation and inference. Proofs of the theoretical results are provided in Appendix.

2. Estimation by linear smoothing

2·1. *Preliminary*

Since linear smoothing procedures have been proved effective for nonparametric curve estimation with independent cross-sectional data, it is natural to extend these methods to the estimation of $\beta(t)$ for observations from longitudinal studies. Here we first give a useful alternative representation of $\beta(t)$, and then present our nonparametric estimators and a crossvalidation criterion for the selection of smoothing parameters.

If we assume that $X(t)$ and $\varepsilon(t)$ are independent, the varying coefficient model (1·2) is

$$Y(t) = X^{\mathrm{T}}(t)\beta(t) + \varepsilon(t). \tag{2·1}$$

Let $E_{XX^{\mathrm{T}}}(t) = E\{X(t)X^{\mathrm{T}}(t)\}$. If we multiply both sides of (2·1) by $X(t)$ and take expectations, it follows that, if $E_{XX^{\mathrm{T}}}(t)$ is invertible, then $\beta(t)$ is unique and given by

$$\beta(t) = E^{-1}_{XX^{\mathrm{T}}}(t)E\{X(t)Y(t)\}. \tag{2·2}$$

If $E_{XX^{\mathrm{T}}}(t)$ does not have a unique inverse, then $\beta(t)$ is not unique and (2·1) becomes unidentifiable. We assume for the rest of the paper that $E^{-1}_{XX^{\mathrm{T}}}(t)$ exists. It is then easy to show that $\beta(t)$ as given in (2·2) uniquely minimises the second moment $E[\{Y(t) - X^{\mathrm{T}}(t)\beta(t)\}^2]$ for any given $t \in R$.

2·2. *Smoothing splines*

Statistical properties and practical implementation of spline methods can be found in Eubank (1988) among others. Suppose that the functions $\beta_0(t), \ldots, \beta_k(t)$ of (1·2) are twice continuously differentiable and their second derivatives $\beta_0''(t), \ldots, \beta_k''(t)$ are bounded and square integrable. A smoothing spline estimator of $\beta_0(t), \ldots, \beta_k(t)$ minimises

$$J(\beta, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[Y_{ij} - \left\{ \sum_{l=0}^{k} X_{ijl}\beta_l(t_{ij}) \right\} \right]^2 + \sum_{l=0}^{k} \lambda_l \int \{\beta_l''(t)\}^2 \, dt, \tag{2·3}$$

where $\lambda = (\lambda_0, \ldots, \lambda_k)^{\mathrm{T}}$ are positive-valued smoothing parameters which penalise the roughness of $\beta_0(t), \ldots, \beta_k(t)$. As in univariate smoothing (Eubank, 1988, pp. 191–207), it can be shown that the minimisers of (2·3) are natural cubic splines with knots located at the distinct values of t_{ij}.

For minimising $J(\beta, \lambda)$ of (2·3), it is convenient to represent $\beta_0(t), \ldots, \beta_k(t)$ in terms of spline basis functions such as B-splines with knots as above. We express each $\beta_l(t)$ in the form

$$\beta_l(t) = \sum_{r=1}^{d} \gamma_{rl} B_r(t) = B^{\mathrm{T}}(t)\gamma_l, \tag{2·4}$$

where $d \geqslant 1$ $(-\infty < t < \infty)$, $\gamma_l = (\gamma_{1l}, \ldots, \gamma_{dl})^{\mathrm{T}}$ are real-valued coefficients, and $B(t) = (B_1(t), \ldots, B_d(t))^{\mathrm{T}}$ is a set of basis functions. We can then find the coefficient vectors γ_l

$(l=0, 1, \ldots, k)$ which minimise the quadratic functional $J(\beta, \lambda)$. For each subject i, let

$$Y_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}, \quad X_{i.l} = \mathrm{diag}(X_{i1l}, \ldots, X_{in_il}), \quad t_i = (t_{i1}, \ldots, t_{in_i})^{\mathrm{T}},$$

$$b_l(t_i) = (\beta_l(t_{i1}), \ldots, \beta_l(t_{in_i}))^{\mathrm{T}}, \quad B_i = \begin{pmatrix} B_1(t_{i1}) & \ldots & B_d(t_{i1}) \\ \vdots & \vdots & \vdots \\ B_1(t_{in_i}) & \ldots & B_d(t_{in_i}) \end{pmatrix}$$

and let Ω be the $d \times d$ matrix whose (i, j)th element is $\Omega_{ij} = \int \{B_i''(t) B_j''(t)\}\, dt$. Then (2·3) is equivalent to

$$J(\beta, \lambda) = \sum_{i=1}^{n} \left\{ Y_i - \sum_{l=0}^{k} (X_{i.l} B_i \gamma_l) \right\}^{\mathrm{T}} \left\{ Y_i - \sum_{l=0}^{k} (X_{i.l} B_i \gamma_l) \right\} + \sum_{l=0}^{k} (\lambda_l \gamma_l^{\mathrm{T}} \Omega \gamma_l). \qquad (2\cdot5)$$

If we set each $\partial J(\beta, \lambda)/\partial \gamma_{rl} = 0$, the minimiser $(\gamma_0, \ldots, \gamma_k)$ of (2·5) satisfies the normal equations

$$\sum_{i=1}^{n} \left\{ (X_{i.l} B_i)^{\mathrm{T}} \sum_{l=0}^{k} (X_{i.l} B_i \gamma_l) \right\} + \sum_{l=0}^{k} (\lambda_l \Omega \gamma_l) = \sum_{i=1}^{n} \{(X_{i.l} B_i)^{\mathrm{T}} Y_i\} \quad (l = 0, \ldots, k). \qquad (2\cdot6)$$

If the normal equations (2·6) have a unique solution $(\hat\gamma_0, \ldots, \hat\gamma_k)$, without loss of generality there exist $d \times n_i$ matrices N_{il}, for $i = 1, \ldots, n$ and $l = 0, \ldots, k$, so that

$$\hat\gamma_l = \sum_{i=1}^{n} (N_{il} Y_i) \quad (l = 0, \ldots, k). \qquad (2\cdot7)$$

The corresponding linear estimators $\hat\beta_0(t), \ldots, \hat\beta_k(t)$ are obtained by substituting $(\gamma_0, \ldots, \gamma_k)$ with $(\hat\gamma_0, \ldots, \hat\gamma_k)$ in (2·4), i.e.

$$\hat\beta_l(t) = \sum_{r=1}^{d} \{\hat\gamma_{rl} B_r(t)\} = \sum_{i=1}^{n} \{B^{\mathrm{T}}(t) N_{il} Y_i\} \quad (l = 0, \ldots, k). \qquad (2\cdot8)$$

The existence and uniqueness of the solution $(\hat\gamma_0, \ldots, \hat\gamma_k)$ of this linear system depend on the design matrices $X_{i.l}$, t_i for $i = 1, \ldots, n$. For practical implementation of smoothing splines, one has to select an adequate smoothing parameter vector λ and basis functions. It can be seen from (2·3) that too large a λ_l gives an excessive penalty for the roughness of β_l, thus resulting in an oversmoothed estimator $\hat\beta_l$. Conversely, too small a λ_l results in an undersmoothed $\hat\beta_l$.

In practice, if a unique solution of (2·6) exists, it may be found directly or by using the backfitting algorithm suggested by Hastie & Tibishirani (1990, p. 91). We note that (2·6) comprises a system of equations of order $(k+1)d \times (k+1)d$, the solution of which can be used to find the estimators for all t. A practical difficulty is that d can be quite large since, as noted above, it is of the order of the number of distinct t_{ij}. Backfitting is one way of coping with this difficulty. It can also be quite adequately circumvented by approximating the smoothing spline solutions by splines with a relatively small number of fixed equispaced knots as in Parker & Rice (1985), thus drastically reducing the dimensionality of the computations. Clearly, deeper theoretical properties of $\hat\beta_0(t), \ldots, \hat\beta_k(t)$ with longitudinal observations deserve further investigation, but are beyond the scope of this paper.

2·3. *Local polynomials*

This class of estimators is a generalisation of kernel type estimators, for which theory and applications with independent cross-sectional data have been studied by Stone (1977), Cleveland (1979), Buja, Hastie & Tibshirani (1989), Hastie & Tibshirani (1990, pp. 29–31) and Fan (1993) among others. This generalisation has many advantages over the kernel methods, particularly in estimation at boundary points; compare Hastie & Loader (1993). Theoretical and simulation results indicate that smoothings with locally weighted polynomials are effective alternatives to smoothing splines.

Motivated by their performance in cross-sectional data, we propose a class of smoothing methods which extends the existing approaches of locally weighted polynomials to longitudinal data. Let $W_{ij}(t)$, for $i=1,\ldots,n$ and $j=1,\ldots,n_i$, be weight functions of t_{ij} and t. In particular, $W_{ij}(t)$ may be selected as a kernel function $K\{(t-t_{ij})h^{-1}\}$, or based on nearest neighbours as in Hastie & Tibshirani (1990, § 2.11). For each $1 \leqslant i \leqslant n$, let $\mathscr{B}_i$ be the $n_i \times d$ basis matrix whose (q,r)th element is $(t_{iq}-t)^{r-1}$ and let $\mathscr{W}_i = \mathrm{diag}(W_{i1}(t),\ldots,W_{in_i}(t))$ be the diagonal weight matrix. Our local polynomial fit $\hat{b}(t) = (\hat{b}_0(t),\ldots,\hat{b}_k(t))$ with $\hat{b}_l(t) = (\hat{b}_{1l}(t),\ldots,\hat{b}_{dl}(t))^{\mathrm{T}}$ minimises the following locally weighted sum of squares:

$$L(t) = \sum_{i=1}^{n}\left[Y_i - \sum_{l=0}^{k}\{X_{i.l}\mathscr{B}_i b_l(t)\}\right]^{\mathrm{T}} \mathscr{W}_i(t)\left[Y_i - \sum_{l=0}^{k}\{X_{i.l}\mathscr{B}_i b_l(t)\}\right]. \tag{2·9}$$

To simplify the notation, let $\mathscr{M}_{il}(t) = (X_{i.l}\mathscr{B}_i)^{\mathrm{T}}\mathscr{W}_i(t)$ and let $\mathscr{N}(t)$ be the matrix whose (r,l)th block with $r,l=0,\ldots,k$ is $\sum_{i=1}^{n}\{(X_{i.l}\mathscr{B}_i)^{\mathrm{T}}\mathscr{W}_i(t)(X_{i.r}\mathscr{B}_i)\}$. Then the normal equations corresponding to (2·9) are

$$\mathscr{N}(t)\hat{b}(t) = \mathscr{M}\circ Y \tag{2·10}$$

where $\mathscr{M}\circ Y$ is the $d+k+1$ column matrix $(\sum_{i=1}^{n}\mathscr{M}_{i0}Y_i,\ldots,\sum_{i=1}^{n}\mathscr{M}_{ik}Y_i)^{\mathrm{T}}$. In order to compute the estimator, if $\mathscr{N}(t)$ is nonsingular, a system of size $d(k+1)\times d(k+1)$ must be solved for each t.

Remark 2·1. As a local constant fit, a kernel estimator is a special case of (2·10) with $d=1$. Denote by X_i the covariate matrix of the ith subject,

$$X_i = \begin{pmatrix} X_{i10} & X_{i11} & \cdots & X_{i1k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{in_i0} & X_{in_i1} & \cdots & X_{in_ik}\end{pmatrix},$$

and by $K_i(t)$ the kernel matrix $\mathrm{diag}(K\{(t-t_{i1})h^{-1}\},\ldots,K\{(t-t_{in_i})h^{-1}\})$. It follows from (2·10) that, if $\sum_{i=1}^{n}X_i^{\mathrm{T}}K_i(t)X_i$ is invertible, the kernel estimator of $\beta(t)$ is

$$\hat{\beta}(t) = \left[\sum_{i=1}^{n}\{X_i^{\mathrm{T}}K_i(t)X_i\}\right]^{-1}\left[\sum_{i=1}^{n}\{X_i^{\mathrm{T}}K_i(t)Y_i\}\right]. \tag{2·11}$$

2·4. *Selection of smoothing parameters*

In practice, the smoothing parameters can be selected subjectively by examining scatter plots and the fitted curves. However, especially when more than one covariate is present, it is useful to develop automatic procedures.

Here we focus on a method of crossvalidation for smoothing parameter choice. Following Rice & Silverman (1991), we use a form of crossvalidation in which single subjects are deleted one at a time, rather than single responses, since the latter procedure is unsuitable when there is intra-subject correlation. The main advantage of the method is that it does not rely on specific correlation structures of the data. Under a different model from (1·2), Hart & Wehrly (1993) have derived a consistency result for this crossvalidation method.

Smoothing in the current context could have a number of different objectives, especially when $\beta(t)$ has more than one component. Interest might focus on the risk of a single or several components of $\beta(t)$ either at a fixed t, such as the mean squared error, or globally, such as the mean integrated squared error. For notational simplicity, we denote by λ the smoothing parameters of any linear estimators of this section. Let $N = \sum_{i=1}^{n} n_i$. We consider averaged predictive squared error,

$$\text{ASPE}_\lambda(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} E\,[\{Y_{ij}^* - X_{ij}^{\mathrm{T}} \hat{\beta}(t_{ij})\}^2], \tag{2·12}$$

where Y_{ij}^* is a new observation at (X_{ij}, t_{ij}), that is $Y_{ij}^* = X_{ij}^{\mathrm{T}}\beta(t_{ij}) + \varepsilon_i^*(t_{ij})$ and $\varepsilon_i^*(t)$ is a new realisation of the stochastic process $\varepsilon(t)$.

Let $\hat{\beta}^{(-i)}(t)$ be an estimator of $\beta(t)$ based on any one of the linear smoothing methods described in §§ 2·2 and 2·3 by leaving out all the observed measurements of the ith subject. The crossvalidation average predictive squared error criterion is defined as

$$\text{CV}(\lambda) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \{Y_{ij} - X_{ij}^{\mathrm{T}} \hat{\beta}^{(-i)}(t_{ij})\}^2. \tag{2·13}$$

Then our crossvalidation smoothing parameter, λ_{CV}, is the minimiser of $\text{CV}(\lambda)$.

Remark 2·2. For the local polynomial estimators given in (2·9), minimising $\text{CV}(\lambda)$ should return a single smoothing parameter, the crossvalidated bandwidth h_{CV}, which may not allow us to obtain appropriate fits to all $\beta_0(t), \ldots, \beta_k(t)$ when they satisfy different smoothness conditions. For smoothing splines of § 2·2, the crossvalidation smoothing parameters consist of $\lambda_{0,\mathrm{CV}}, \ldots, \lambda_{k,\mathrm{CV}}$. Intuitively, the extra number of smoothing parameters in smoothing splines can be used to allow for possibly different smoothness of the nonparametric components. Note that in (2·3) the smoothing parameters λ_l have units corresponding to those of the respective covariates. Thus, if it is desired for simplicity to use a single smoothing parameter, the covariates should be standardised.

Remark 2·3. In practice, $\text{CV}(\lambda)$ can only be minimised within a preselected compact set of the parameter space, and there may be more than one local minimum. Thus, it is often useful first to try several λ subjectively, and then determine a workable range of λ by examining the fits and the $\text{CV}(\lambda)$ values in this range. If there is more than one local minimum, it is also helpful to re-examine the fit of $\hat{\beta}$ with each λ which gives the corresponding local minimum, instead of only considering the value of λ which gives the global minimum.

Remark 2·4. The effect of leaving out one single subject in the computation of $\text{CV}(\lambda)$ can be explicitly calculated as a perturbation from the full solution, providing an alternative expression of $\text{CV}(\lambda)$ which is useful to speed up computation. To see how this works for kernel estimators, for example, note first that, by (2·11),

$$A(t)\hat{\beta}(t) = \sum_{i=1}^{n} \{J_i(t)Y_i\}, \quad A^{(-i)}(t)\hat{\beta}^{(-i)}(t) = \sum_{j=1}^{n} \{J_j(t)Y_j\} - J_i(t)Y_i,$$

where

$$J_i(t) = X_i^{\mathrm{T}} K_i(t), \quad A(t) = \sum_{i=1}^{n} \{X_i^{\mathrm{T}} K_i(t) X_i\}, \quad A^{(-i)}(t) = \sum_{j \neq i} \{X_j^{\mathrm{T}} K_j(t) X_j\}.$$

By the well-known matrix updating formula, e.g. Cook & Weisberg (1982, eqn (A.2.1)), we have

$$\{A^{(-i)}(t)\}^{-1} = \{A(t) - X_i^{\mathrm{T}} J_i^{\mathrm{T}}(t)\}^{-1}$$
$$= A^{-1}(t) + A^{-1}(t) X_i^{\mathrm{T}} \{I - J_i^{\mathrm{T}}(t) A^{-1}(t) X_i^{\mathrm{T}}\}^{-1} J_i^{\mathrm{T}}(t) A^{-1}(t), \qquad (2{\cdot}14)$$

where $J_i^{\mathrm{T}}(t) = K_i(t) X_i$ and I is the $n_i \times n_i$ identity matrix. Then (2·11) implies that

$$\mathrm{CV}(h) = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - X_{ij}^{\mathrm{T}} \{A^{(-i)}(t_{ij})\}^{-1} \left[\sum_{i'=1}^{n} \{J_{i'}(t_{ij}) Y_{i'}\} - J_i(t_{ij}) Y_i \right] \right)^2,$$

where $\{A^{(-i)}(t_{ij})\}^{-1}$ can be computed using the right-hand side of (2·14). Similar expressions can also be derived for smoothing splines and the local polynomials.

3. An application to growth of children

The data considered here involve infants' genders and HIV infection status, that is HIV positive or negative, measured one year after birth, the third trimester maternal vitamin A levels during pregnancy and repeatedly measured weights of 328 infants from an African AIDS cohort study at the Johns Hopkins University. The covariates in this example are not time varying, although we allow their coefficients to be; further discussion of this structure is contained in § 5. All infants were born from HIV infected mothers in central Africa and survived beyond one year of age. The follow-up study lasted two years and infants' weights were recorded during every scheduled monthly visit. For various reasons, a number of the scheduled visits were missed by some infants, resulting in unequal numbers of repeated weight measurements per infant. The main objective is to evaluate the time-varying effects of two binary covariates, child's gender and HIV status, and one continuous covariate, the third trimester maternal vitamin A level, on child's weight. Previous studies have shown that vitamin A can improve immune function and resistance to disease (Semba, 1994). Biologically, a significant association between maternal vitamin A levels and infant growth may suggest the benefit of vitamin A supplementation in mother's and infant's diets.

This dataset was initially analysed by Semba et al. (1997), where vitamin A was treated as a binary covariate, i.e. deficiency and nondeficiency, and the growth prediction curves are obtained using parametric regression models and generalised estimation equations (Diggle et al., 1994). Here, for $j = 1, \ldots, n_i$, Y_{ij} is the weight in kilograms of the ith infant at time t_{ij} after birth, $X_{ij0} = 1$, $X_{ij1} = 1$ if the ith infant is HIV positive, and $X_{ij1} = 0$ if he/she is HIV negative, X_{ij2} denotes the ith infant's maternal vitamin A level, and X_{ij3} takes value 0 or 1 according as the ith infant is female or male, respectively. For brevity, we only present the smoothing results of (1·2) based on kernel and spline methods. The smoothing results of the local polynomials are very similar to those given by kernels and smoothing splines.

Figure 1 shows the estimated values of $\beta_l(t)$ $(l = 0, \ldots, 3)$ together with their ± 2 pointwise bootstrap standard error bands. These curves were computed using natural cubic splines with λ_l $(l = 0, \ldots, 3)$, chosen to be the crossvalidated smoothing parameters 0·125, 0·2, 2·5 and 0·2, respectively. The bootstrap standard errors were computed at 100 selected time points using 200 bootstrap replications by resampling from the subjects, i.e. randomly resampling the entire repeated measurements of the subjects with replacement. Since smoothing bias has not been taken into account, these bootstrap bands are not actually confidence intervals in the usual sense. From Fig. 1 it is seen that the magnitudes of the coefficients corresponding to the effects of time, gender and vitamin A initially increase

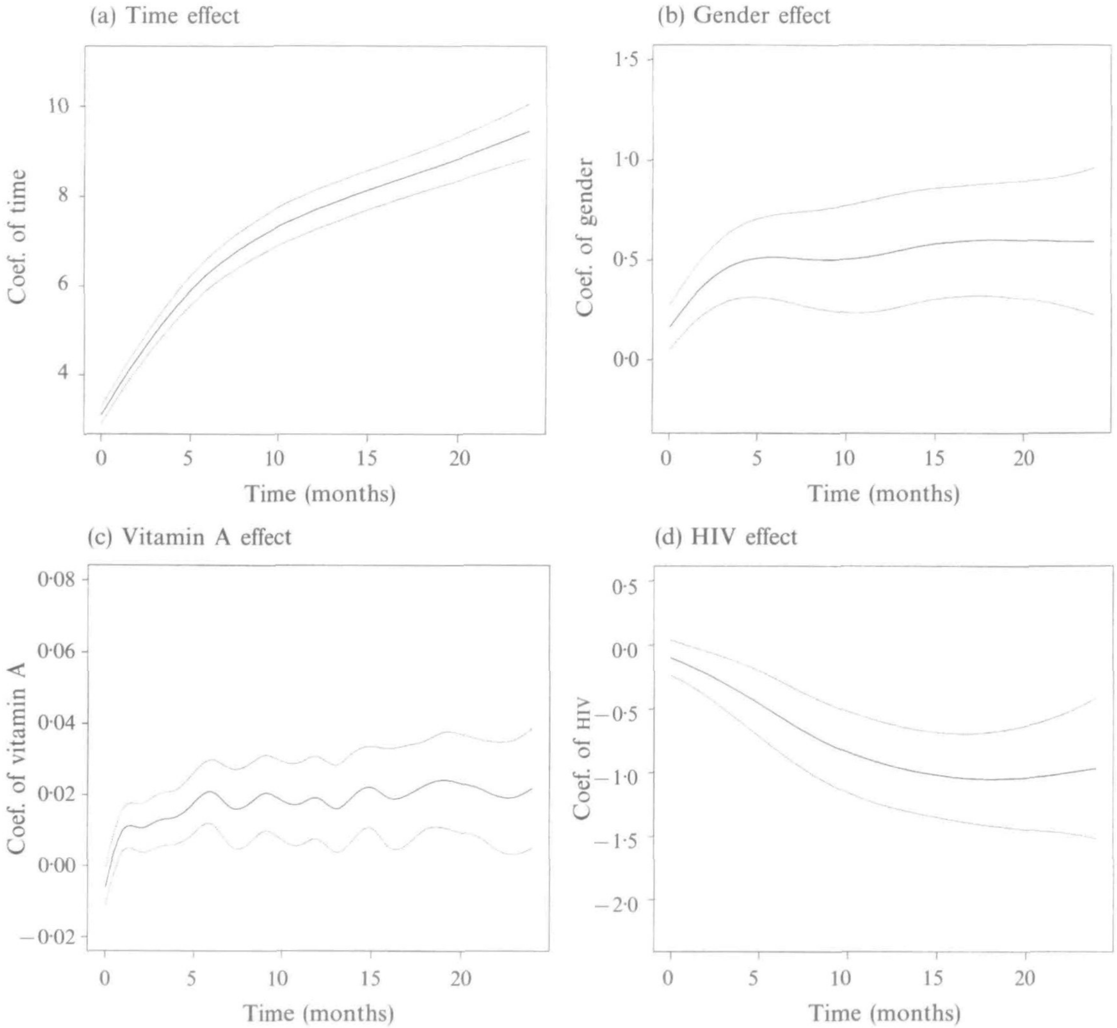

Fig. 1. Estimators using natural cubic splines with λ_l $(l = 0, \ldots, 3)$ chosen to be the crossvalidated smoothing parameters 0·125, 0·2, 2·5 and 0·2, respectively. The dotted curves represent the ± 2 bootstrap standard error bands. (a) Time effect: $\hat{\beta}_0(t)$ versus time. (b) Gender effect: $\hat{\beta}_3(t)$ versus time. (c) Vitamin A effect: the estimated effect of vitamin A $\hat{\beta}_2(t)$ versus time. (d) HIV effect: $\hat{\beta}_1(t)$ versus time.

with time and then level off, while the effect of HIV decreases with time and then levels off. The initial change with time probably reflects the cumulative effects of additional diseases early in life due to HIV infection and/or low vitamin A levels. The levelling off of the difference may be due to the establishment of the infants' immunity function at one year of age and frailty effects from the sickest and lowest weight babies dying. The smoothness of $\hat{\beta}_0(t)$, $\hat{\beta}_1(t)$ and $\hat{\beta}_3(t)$ seem to be adequate, but $\hat{\beta}_2(t)$ appears to be slightly undersmoothed.

Figure 2 shows the estimated values of $\beta_l(t)$ $(l = 0, \ldots, 3)$ and their ± 2 pointwise bootstrap standard error bands computed using (2·11) with the standard Gaussian kernel and the crossvalidated bandwidth $h_{CV} = 0{\cdot}5$. The estimated curves are undersmoothed and the oscillations in the curves reflect the monthly visiting schedule; crossvalidation has apparently not been effective. From Figs 1 and 2 it appears that the smoothing splines are generally superior to the kernel estimators, possibly because of the use of multiple smoothing

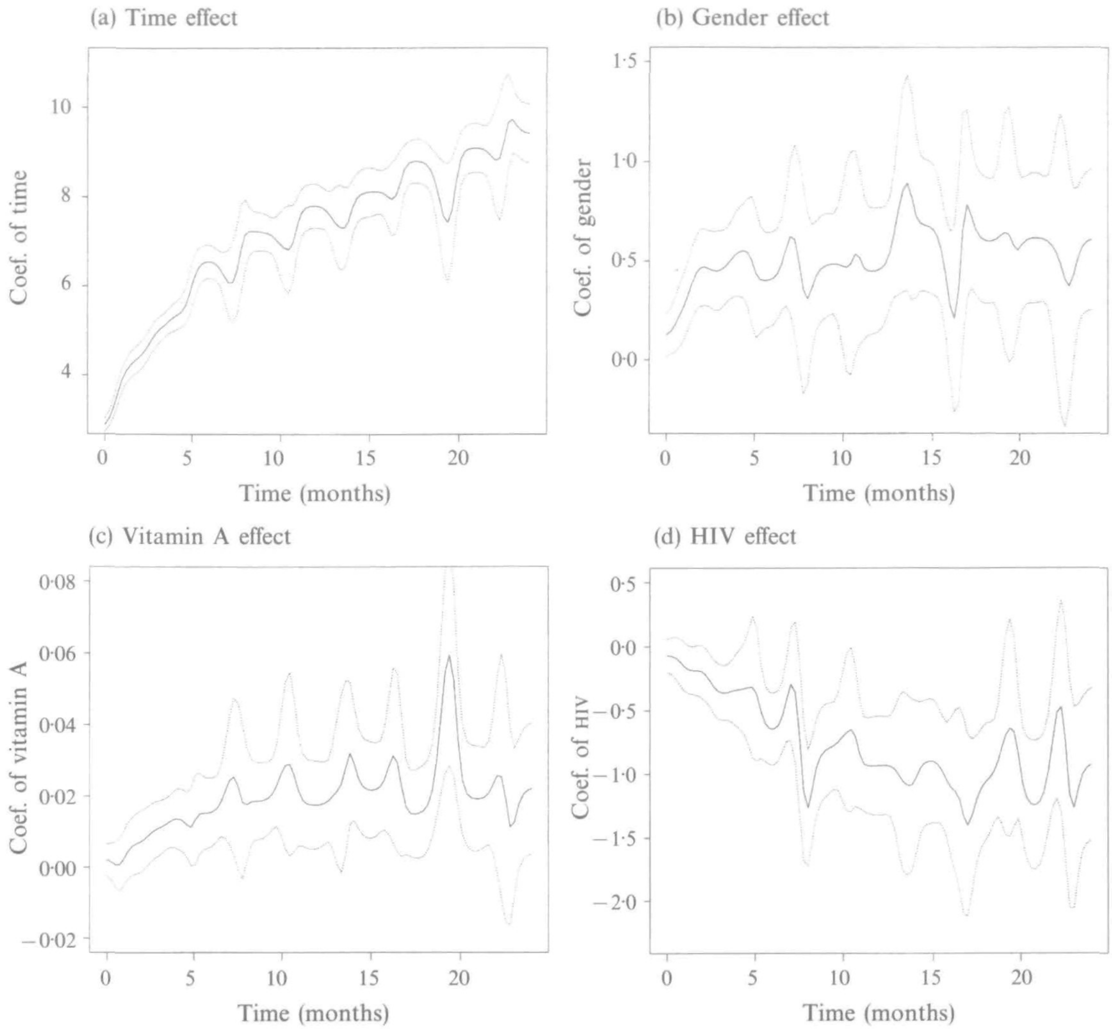

Fig. 2. Estimators using kernel method with the standard Gaussian kernel and the crossvalidated bandwidth $h_{\text{CV}} = 0{\cdot}5$. The dotted curves represent the ± 2 bootstrap standard error bands. (a) Time effect: $\hat{\beta}_0(t)$ versus time. (b) Gender effect: $\hat{\beta}_3(t)$ versus time. (c) Vitamin A effect: the estimated effect of vitamin A $\hat{\beta}_2(t)$ versus time. (d) HIV effect: $\hat{\beta}_1(t)$ versus time.

parameters, but the quantitative trends of the fitted curves are similar in both Fig. 1 and Fig. 2.

4. Asymptotic risks of kernel estimators

4·1. *Notation and mean squared errors*

For simplicity, we only consider in this section the asymptotic risk representations under mean squared error for the kernel estimators as defined in (2·11). The main result of this section has two interesting features, which distinguish the asymptotic risk of $\hat{\beta}(t)$ from that of the Nadaraya–Watson kernel estimator in nonparametric regression. First, the asymptotic bias of our kernel estimator $\hat{\beta}(t)$ is affected by the smoothness of $X(t)$ as well as the smoothness of $\beta(t)$ and the underlying design density of t. Secondly, the asymptotic variance of $\hat{\beta}(t)$ is influenced by the intra-subject correlation of the data and the numbers of repeated measurements n_i, as well as the variance of $\varepsilon(t)$. We believe that the asymptotic risks of smoothing splines and local polynomials may be analogously investigated, but we have not done so.

The estimation methods of § 2 can accommodate both fixed and random designs. For technical convenience, we assume that the design points t_{ij}, for $j=1,\ldots,n_i$ and $i=1,\ldots,n$, are chosen independently according to some design distribution F_T and design density f_T.

Throughout this section we write, for all $i=1,\ldots,n$, $j=1,\ldots,n_i$ and $r, l=0,\ldots,k$,

$$\sigma^2(t)=E\{\varepsilon^2(t)\}, \quad \rho_\varepsilon(t)=\lim_{\delta\to 0} E\{\varepsilon(t+\delta)\varepsilon(t)\}, \quad \xi_{lr}(t)=E(X_{ijl}X_{ijr}\mid t_{ij}=t),$$

and assume that $\beta_l(t)\,f_T(t)$ and $\xi_{lr}(t)$ are twice continuously differentiable, $\sigma^2(t)$ and $\rho_\varepsilon(t)$ are continuous, and $E(X_{ijl}^4)$ is finite. In general, $\rho_\varepsilon(t)$ does not necessarily equal $\sigma^2(t)$; see, for example, Zeger & Diggle (1994). The kernel is assumed to be square integrable, integrate to one and have finite second moment, while the bandwidth satisfies $h\to 0$ and $nh\to\infty$ as $n\to\infty$. These assumptions are comparable with the regularity conditions commonly used in nonparametric regression under independent cross-sectional data, e.g. Härdle (1990, pp. 91–3), and are general enough to be satisfied in many interesting practical situations. Theoretically, these assumptions could be further modified or weakened in various ways so that more desirable asymptotic properties of the kernel estimator $\hat{\beta}(t)$ may be obtained. A generalisation of the results presented in this section can be found in the Appendix.

The risk of $\hat{\beta}(t)$ depends on the choice of loss function. Let w_l ($l=0,\ldots,k$) be nonnegative constants. We consider the following local squared risk of $\hat{\beta}(.)$ at time t,

$$\text{MSE}_w\{\hat{\beta}(t)\}=E[\{\hat{\beta}(t)-\beta(t)\}^{\mathrm{T}}w\{\hat{\beta}(t)-\beta(t)\}], \tag{4·1}$$

where $w=\operatorname{diag}(w_0,\ldots,w_k)$ with nonnegative diagonal elements w_l. Unfortunately, a minor technical inconvenience for the kernel estimator $\hat{\beta}(t)$ is that its mean squared error as defined in (4·1) may not exist in general; see Rosenblatt (1969) and an unpublished North Carolina Institute of Statistics report by W. Härdle and J. S. Marron. Thus, a slight modification of (4·1) has to be considered. By straightforward algebra, we have the approximation

$$\Delta(t)=\{1+o_p(1)\}\{\hat{\beta}(t)-\beta(t)\}=f_T^{-1}(t)\{E_{XX^{\mathrm{T}}}(t)\}^{-1}\hat{R}(t), \tag{4·2}$$

where

$$\hat{R}(t)=\left\{(Nh)^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}K_i(t)Y_i\right\}-\left\{(Nh)^{-1}\sum_{i=1}^{n}X_i^{\mathrm{T}}K_i(t)X_i\right\}\beta(t).$$

To avoid the technical inconvenience that might arise because of nonexistence of the mean squared errors, the asymptotic risk of $\hat{\beta}(t)$ is described through the modified mean squared error

$$\text{MSE}_w^*\{\hat{\beta}(t)\}=E\{\Delta(t)^{\mathrm{T}}w\Delta(t)\}=\sum_{l=0}^{k}\sum_{r=0}^{k}[M_{lr}(t)E\{\hat{R}_l(t)\hat{R}_r(t)\}],$$

where $\hat{R}_l(t)$ is the lth element of the $k+1$ column vector $\hat{R}(t)$ and $M_{lr}(t)$ is the (l,r)th element of the $(k+1)\times(k+1)$ matrix $M(t)=f_T^{-2}(t)[\{E_{XX^{\mathrm{T}}}(t)\}^{-1}]^{\mathrm{T}}w\{E_{XX^{\mathrm{T}}}(t)\}^{-1}$. We measure the bias of $\hat{\beta}(t)$ by $B^*\{\hat{\beta}(t)\}=E\{\Delta(t)\}=f_T^{-1}(t)\{E_{XX^{\mathrm{T}}}(t)\}^{-1}E\{\hat{R}(t)\}$, and the variance of $\hat{\beta}_l(t)$ by $V^*\{\hat{\beta}_l(t)\} = \text{var}\{\Delta_l(t)\}$, with $\Delta_l(t)$ being the lth component of $\Delta(t)$. Let the lth component of $B^*\{\hat{\beta}(t)\}$ be denoted by $B^*\{\hat{\beta}_l(t)\}$. As in the usual nonparametric regression setting, we have the following variance-bias squared decomposition

$$\text{MSE}^*_w\{\hat{\beta}(t)\} = \sum_{l=0}^{k} w_l [B^*\{\hat{\beta}_l(t)\}]^2 + \sum_{l=0}^{k} w_l V^*\{\hat{\beta}_l(t)\}. \tag{4·3}$$

4·2. *Asymptotic risk representations*

Define

$$b_l(t) := \sum_{r=0}^{k} \sum_{a=0}^{1} \sum_{b=0}^{a} \left[\frac{\beta_r^{(2-a)}(t)\xi_{lr}^{(a-b)}(t) f_T^{(b)}(t)}{(2-a)!(a-b)!b!} \left\{ \int u^2 K(u)\, du \right\} \right],$$

$$Z_1(t) := \rho_\varepsilon(t) \left[\sum_{l_1=0}^{k} \sum_{l_2=0}^{k} \{M_{l_1 l_2}(t) \xi_{l_1 l_2}(t)\} \right], \quad Z_2(t) := \sigma^2(t) \left[\sum_{l_1=0}^{k} \sum_{l_2=0}^{k} \{M_{l_1 l_2}(t) \xi_{l_1 l_2}(t)\} \right].$$

It can be derived from the general results of the Appendix that, when n is sufficiently large, the bias and variance terms of (4·3) are, respectively,

$$B^*\{\hat{\beta}(t)\} = f_T^{-1}(t)\{E_{XX^{\mathrm{T}}}(t)\}^{-1} h^2 (b_0(t), \ldots, b_k(t))^{\mathrm{T}} + o(h^2), \tag{4·4}$$

$$\sum_{l=0}^{k} w_l V^*\{\hat{\beta}_l(t)\} = \left\{ \sum_{i=1}^{n} \left(\frac{n_i}{N} \right)^2 \right\} f_T^2(t) Z_1(t) + \frac{1}{Nh} f_T(t) \left\{ \int K^2(u)\, du \right\} Z_2(t)$$
$$+ o\left\{ \frac{1}{Nh} + \sum_{i=1}^{n} \left(\frac{n_i}{N} \right)^2 \right\}. \tag{4·5}$$

Thus, the asymptotic representation of $\text{MSE}^*_w\{\hat{\beta}(t)\}$ is obtained by substituting the right-hand sides of (4·4) and (4·5) into (4·3). Since $h \to 0$ and $Nh \to \infty$ as $n \to \infty$, $\hat{\beta}(t)$ is consistent, that is $\text{MSE}^*_w\{\hat{\beta}(t)\} \to 0$ as $n \to \infty$, when $\sum_{i=1}^{n} (n_i N^{-1})^2 \to 0$, which holds if and only if $\max_{1 \leqslant i \leqslant n} (n_i N^{-1}) \to 0$ as $n \to \infty$.

Remark 4·1. It is interesting to see from (4·5) that the intra-subject correlations of longitudinal data are only included in $Z_1(t)$ in the variance term (4·5). Without this extra term, the asymptotic behaviours of $\hat{\beta}(t)$ would be the same as with independent cross-sectional samples. If n_i are bounded and n is sufficiently large, then the asymptotic variance term is dominated by the second term of the right-hand side of (4·5) and, if we minimise the dominating terms of the asymptotic $\text{MSE}^*_w\{\hat{\beta}(t)\}$, the optimal bandwidth is $h_{\text{opt}} = O(N^{-1/5})$. The consequent mean squared error $\text{MSE}^*_w\{\hat{\beta}_{h_{\text{opt}}}(t)\}$ is of the order $N^{4/5}$.

Remark 4·2. In general, $\hat{\beta}(t)$ is not necessarily a consistent estimator of $\beta(t)$ when only N converges to infinity. For example, if $n_i = m$ $(i = 1, \ldots, n)$, m converges to infinity but n stays bounded, then, since $N^{-2}(\sum_{i=1}^{n} n_i^2 - N) = n^{-1} - N^{-1}$ is bounded away from zero for sufficiently large N, $\text{MSE}^*_w\{\hat{\beta}(t)\}$ does not converge to zero as N goes to infinity.

Remark 4·3. Asymptotically, the effect of the intra-subject correlations on $\text{MSE}^*_w\{\hat{\beta}(t)\}$ only depends on the limiting values, $\rho_\varepsilon(t)$, of the covariances of $\varepsilon_i(t)$ and $\varepsilon_i(s)$ as $s \to t$. This is because of the local averaging nature of kernel methods: the estimators tend to ignore the measurements at design points t_{ij} which are outside a shrinking neighbourhood of t. Since the bandwidths shrink to zero, any correlation between $\varepsilon_i(t)$ and $\varepsilon_i(s)$ $(t \neq s)$ is ignored when n is sufficiently large. This is fortunate, since in practice we may only be aware of the presence of the intra-subject correlations but have very little knowledge about the specific correlation structures. By using a local smoothing method, we essentially choose to ignore the correlation structures. One needs to bear in mind, however, that asymptotic results on nonparametric regression should be generally viewed with caution and that this is especially true in the current complicated context. These results may only provide some qualitative insight for the adequacy of the estimation procedures.

5. DISCUSSION

We have demonstrated the utility of linear smoothing for the varying coefficient model (1·2). Since they involve multiple smoothing parameters, smoothing splines have advantages over local polynomials. Many practical and theoretical issues deserve further study, especially the following.

Time independent covariates. In many situations such as the epidemiological study of § 3, the covariates X do not depend on t, and only the outcome variable Y is repeatedly measured. Then, if we assume $E(XX^{\mathrm{T}})$ is invertible, (2·2) reduces to $\beta(t) = E^{-1}(XX^{\mathrm{T}})E\{XY(t)\}$. An obvious estimator of $E(XX^{\mathrm{T}})$ is its corresponding sample average. Thus, computationally simpler estimators of $\beta(t)$ may be developed by forming a smooth estimator of $E\{XY(t)\}$ only.

Inference. We have used bootstrap standard errors to assess variability, but we have not addressed some other important inferential issues. Various types of confidence region might be desired, such as intervals for components or linear combinations of components of $\beta(t)$ for fixed t and simultaneous confidence bands for all t in an interval. Various hypothesis testing problems are of interest as well; for example, we may wish to test that a certain component of $\beta(.)$ is identically zero or constant. The bootstrap provides a natural approach to such problems, but the theoretical and practical aspects would require substantial development.

ACKNOWLEDGEMENT

This research was partially supported by grants from the National Institute on Drug Abuse, the National Institutes of Health and the Office of Nutrition, Bureau for Science and Technology. Partial support for John Rice was also provided by the National Science Foundation. The authors are grateful to Professor Richard D. Semba for providing his valuable data on children's growth and vitamin A, to Professor Shih-Ping Han for insightful discussions and comments on the properties of splines and to Mr. Chin-Tsang Chiang for many numerical computations used in this paper. We would also like to thank three referees for many insightful comments and suggestions which greatly improved the presentation of this paper.

APPENDIX

Proofs of the main results

We state and outline the proof of a general result on the mean squared risk of the kernel estimator $\hat{\beta}(t)$. To generalise the assumptions of § 4·1, we assume here that (i) $\xi_{lr}(t)$ is Lipschitz continuous with order α_0, that is $|\xi_{lr}(s_1) - \xi_{lr}(s_2)| \leqslant c_0|s_1 - s_2|^{\alpha_0}$ for s_1 and s_2 in the support of f_T and some $c_0 > 0$, $\beta_l(t)$ and $f_T(t)$ are Lipschitz continuous, respectively, with orders $\alpha_1 > 0$ and $\alpha_2 > 0$, (ii) $\sigma^2(t)$

and $\rho_\varepsilon(t)$ are continuous, and (iii) the kernel K is square integrable, integrates to one and satisfies $\int u^\alpha K(u)\,du < \infty$ with $\alpha = \max(\alpha_0, \alpha_1, \alpha_2)$, while $h \to 0$ and $Nh \to \infty$ as $n \to \infty$.

THEOREM 1. *When the number of subjects n is large, t is an interior point of the support of f_T and the above conditions are satisfied, the variance term* (4·5) *still holds, while the bias of $\hat\beta(t)$ of* (2·11) *is given by*

$$B^*\{\hat\beta(t)\} = f_T^{-1}(t)\{E_{XX^{\mathrm{T}}}(t)\}^{-1}(B_0(t), \ldots, B_k(t))^{\mathrm{T}}, \tag{A·1}$$

where, for $l = 0, \ldots, k$,

$$B_l(t) = \sum_{r=0}^{k} \int \{\beta_r(t-hu) - \beta_r(t)\}\xi_{lr}(t-hu) f_T(t-hu) K(u)\,du.$$

The asymptotic $\mathrm{MSE}_w^*\{\hat\beta(t)\}$ *is obtained by substituting* (A·1) *and* (4·5) *into* (4·3). *Furthermore,* $\mathrm{MSE}_w^*\{\hat\beta(t)\} \to 0$ *if and only if* $\max_{1 \leq i \leq n}(n_i N^{-1}) \to 0$ *as* $n \to \infty$.

Proof. Following the approximation of (4·2), it suffices to study the asymptotic representations of $E\{\hat R(t)\}$ and $E\{\hat R_l(t)\hat R_r(t)\}$ for $l, r = 0, \ldots, k$. Define

$$a_{ijl}(t) = \sum_{s=0}^{k} [X_{ijl}X_{ijs}\{\beta_s(t_{ij}) - \beta_s(t)\}] + X_{ijl}\varepsilon_i(t_{ij}).$$

It can be verified by direct computation that

$$\hat R_l(t) = N^{-1}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{a_{ijl}(t)K\left(\frac{t-t_{ij}}{h}\right)\right\}, \tag{A·2}$$

and, since $E\{a_{ijl}(t) \mid t_{ij} = s\} = \sum_{r=0}^{k}\{\beta_r(s) - \beta_r(t)\}\xi_{lr}(s)$,

$$E\{\hat R_l(t)\} = N^{-1}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\int E\{a_{ijl}(t) \mid t_{ij} = s\}K\left(\frac{t-s}{h}\right)f_T(s)\,ds = B_l(t). \tag{A·3}$$

Thus (A·1) follows from (A·2) and (A·3).

To derive the variance term (4·5), we consider the decomposition $\hat R_l(t)\hat R_r(t) + A_{lr1} + A_{lr2} + A_{lr3}$, where

$$A_{lr1} = N^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{a_{ijl}(t)a_{ijr}(t)K^2\left(\frac{t-t_{ij}}{h}\right)\right\},$$

$$A_{lr2} = N^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j \neq j'}\left\{a_{ijl}(t)a_{ij'r}(t)K\left(\frac{t-t_{ij}}{h}\right)K\left(\frac{t-t_{ij'}}{h}\right)\right\},$$

$$A_{lr3} = N^{-2}h^{-2}\sum_{i \neq i'}\sum_{j,j'}\left\{a_{ijl}(t)a_{i'j'r}(t)K\left(\frac{t-t_{ij}}{h}\right)K\left(\frac{t-t_{i'j'}}{h}\right)\right\}.$$

By direct calculations and the change of variables, it is straightforward to verify that

$$\sum_{l=0}^{k}\sum_{r=0}^{k}\{M_{lr}(t)E(A_{lr1})\} = (Nh)^{-1}f_T(t)\left\{\int K^2(u)\,du\right\}Z_2(t) + o(N^{-1}h^{-1}).$$

Using the Cauchy–Schwarz inequality, we can show that

$$\sum_{l=0}^{k}\sum_{r=0}^{k}\{M_{lr}(t)E(A_{lr2})\} = N^{-2}\left(\sum_{i=1}^{n} n_i^2 - N\right)f_T^2(t)Z_1(t) + o\left(N^{-2}\sum_{i=1}^{n} n_i^2\right).$$

Finally, we can verify directly that

$$\sum_{l=0}^{k} w_l V^*\{\hat\beta_l(t)\} = \sum_{l=0}^{k}\sum_{r=0}^{k}\{M_{lr}(t)E(A_{lr1} + A_{lr2})\}.$$

Consequently, (4·5) holds.

Suppose $\sum_{l,r=0}^{k} M_{lr}(t)\xi_{lr}(t) > 0$. Then $\text{MSE}_w^*\{\hat{\beta}(t)\} \to 0$ if and only if $N^{-2}\sum_{i=1}^{n} n_i^2 \to 0$ as $n \to \infty$. It is easy to see that $\sum_{i=1}^{n}(n_i N^{-1})^2 \to 0$ implies $\max_{1 \leqslant i \leqslant n}(n_i N^{-1}) \to 0$. It suffices to show that $\max_{1 \leqslant i \leqslant n}(n_i N^{-1}) \to 0$ implies $\sum_{i=1}^{n}(n_i N^{-1})^2 \to 0$. Assume now that $\max_{1 \leqslant i \leqslant n}(n_i N^{-1}) \to 0$. Then, for any $\varepsilon > 0$, $\max_{1 \leqslant i \leqslant n}(n_i N^{-1}) < \frac{1}{2}\varepsilon$ for sufficiently large n. Let $1 = k_0 < k_1 < \ldots < k_m = n$ be positive integers such that $\frac{1}{2}\varepsilon < \sum_{i=k_{l-1}}^{k_l}(n_i N^{-1}) < \varepsilon$ for $l = 1, \ldots, m-1$, and $\sum_{i=k_{m-1}}^{k_m}(n_i N^{-1}) < \varepsilon$. Then, for all $l = 1, \ldots, m$, $\sum_{i=k_{l-1}}^{k_l}(n_i N^{-1})^2 < \varepsilon^2$. Since $N = \sum_{i=1}^{n} n_i$, we must have $m \leqslant 2/\varepsilon$ and, consequently, $\sum_{i=1}^{n}(n_i N^{-1})^2 < 2\varepsilon$. Since ε can be arbitrarily small, this inequality implies that $\lim_{n\to\infty}\sum_{i=1}^{n}(n_i N^{-1})^2 = 0$. □

REFERENCES

ALTMAN, N. S. (1990). Kernel smoothing of data with correlated errors. *J. Am. Statist. Assoc.* **85**, 749–59.

BUJA, A., HASTIE, T. J. & TIBSHIRANI, R. J. (1989). Linear smoothers and additive models (with Discussion). *Ann. Statist.* **17**, 453–555.

CLEVELAND, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829–36.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

DIGGLE, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959–71.

DIGGLE, P. J., LIANG, K. Y. & ZEGER, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge U.K.: Cambridge University Press.

HART, J. D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B **53**, 173–87.

HART, J. D. & WEHRLY, T. E. (1986). Kernel regression estimation using repeated measurements data. *J. Am. Statist. Assoc.* **81**, 1080–8.

HART, J. D. & WEHRLY, T. E. (1993). Consistency of cross-validation when the data are curves. *Stoch. Proces. Applic.* **45**, 351–61.

HASTIE, T. J. & LOADER, C. (1993). Local regression: automatic kernel carpentry (with Discussion). *Statist. Sci.* **8**, 120–43.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

HASTIE, T. J. & TIBSHIRANI, R. J. (1993). Varying-coefficient models (with Discussion). *J. R. Statist. Soc.* B **55**, 757–96.

JONES, R. H. & ACKERSON, L. M. (1990). Serial correlation in unequally spaced longitudinal data. *Biometrika* **77**, 721–31.

JONES, R.H. & BOADI-BOTENG, F. (1991). Unequally spaced longitudinal data with serial correlation. *Biometrics* **47**, 161–75.

MOYEED, R. A. & DIGGLE, P. J. (1994). Rates of convergence in semi-parametric modeling of longitudinal data. *Aust. J. Statist.* **36**, 75–93.

PANTULA, S. G. & POLLOCK, K. H. (1985). Nested analysis of variance with autocorrelated errors. *Biometrics* **41**, 909–20.

PARKER, R. L. & RICE, J. (1985). Comment on paper by B. W. Silverman. *J. R. Statist. Soc.* B **47**, 40–2.

RICE, J. A. & SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B **53**, 233–43.

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis* **2**, Ed. P. R. Krishnaiah, pp. 25–31. Amsterdam: North-Holland.

SEMBA, R. D. (1994). Vitamin A immunity and infection. *Clin. Inf. Dis.* **19**, 489–99.

SEMBA, R. D., MIOTTI, P., CHIPHANGWI, J. D., HENDERSON, R., DALLABETTA, G., YANG, L. P. & HOOVER, D. R. (1997). Maternal vitamin A deficiency and child growth failure during human immunodeficiency virus infection. *J. Acq. Immun. Defic. Synd.* **14**, 219–22.

STONE, C. J. (1997). Consistent nonparametric regression (with Discussion). *Ann. Statist.* **5**, 549–645.

WARE J. H. (1985). Linear models for the analysis of longitudinal studies. *Am. Statistician* **39**, 95–101.

ZEGER, S. L. & DIGGLE, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–99.

[*Received April* 1996. *Revised July* 1997]