Title
Generalized Varying-coefficient models

Authors
Zong-wu Cai
Jianqing Fan
Run-Ze Li

Publication Date
2011-10-25

Generalized Varying-Coefficient Models

Zongwu Cai
Department of Mathematics
University of North Carolina
Charlotte, NC 28223

Jianqing Fan *
Department of Statistics
University of California
Los Angeles, CA 90095

Run-Ze Li
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

Abstract

This paper deals with statistical inferences based on the generalized varying-coefficient models proposed by Hastie and Tibshirani (1993). Local polynomial regression techniques are used to estimate coefficient functions and the asymptotic normality of the resulting estimators is established. The standard error formulas for estimated coefficients are derived and are empirically tested. A goodness-of-fit test technique, based on a nonparametric maximum likelihood ratio type of test, is also proposed to detect whether certain coefficient functions in a varying-coefficient model are constant or whether any covariates are statistically significant in the model. The null distribution of the test is estimated by a conditional bootstrap method. Our estimation techniques involve solving hundreds of local likelihood equations. To reduce computation burden, a one-step Newton-Raphson estimator is proposed and implemented. We show that the resulting one-step procedure can save computational cost in an order of tens without deteriorating its performance, both asymptotically and empirically. Both simulated and real data examples are used to illustrate our proposed methodology.

Key Words: Asymptotic normality; bootstrap; generalized linear models; goodness-of-fit; local polynomial fitting; one-step procedure.

*Partially supported by NSF Grant DMS-9803200 and NSA 96-1-0015.

1 Introduction

Generalized linear models are widely used in many statistical applications. They are based on two fundamental assumptions: the conditional distributions belong to an exponential family and a known transform of the underlying regression function is linear. Various attempts have been made to relax the above model assumptions and hence widen their applicability, since a wrong model on the regression function can lead to excessive modeling biases and hence erroneous conclusions. For example, generalized additive models (Hastie and Tibshirani 1990) extend traditional linear assumptions by allowing nonparametric additive contributions to a known transform of the regression function, and generalized varying-coefficient models (Hastie and Tibshirani 1993) widen the scope of applications by allowing regression coefficients to depend on certain covariates.

A motivation of this study comes from an analysis of an environmental data set, collected in Hong Kong from January 1, 1994 to December 31, 1995 (Courtesy of Professor T. S. Lau), which consists of weekly measurements of pollutants and other environmental factors. Of interest is to examine the association between the levels of pollutants and the number of weekly total hospital admissions for circulationary and respirationary problems. It is natural to allow the association to change over time (see Figure 3(a) below). Such a problem can be modeled as follows

$$g\{m(\mathbf{u},\, \mathbf{x})\} = \sum_{j=1}^{p} a_j(\mathbf{u})\, x_j \tag{1.1}$$

for some given link function $g(\cdot)$, where $\mathbf{x} = (x_1, \ldots, x_p)^T$, and $m(\mathbf{u}, \mathbf{x})$ is the mean regression function of the response variable Y given the covariates $\mathbf{U} = \mathbf{u}$ and $\mathbf{X} = \mathbf{x}$. For the aforementioned example, the log-link is used, $\mathbf{U}$ is the time covariate, and $\mathbf{X}$ denotes the levels of pollutants. The conditional distribution of the number of weekly hospital admissions given the covariates can be modeled reasonably as a Poisson distribution with the mean function given by (1.1). This is an example of generalized varying-coefficient models. In another context, one is interested in studying how the variables such as burn area and gender affect survival probabilities for different age of burn victims. This is another example of the generalized varying-coefficient models. Detailed analyses of these two data sets will be reported in §4.

By regarding $X_1 \equiv 1$, (1.1) permits varying intercept term in the model. In particular, when the coefficient functions $a_j(\cdot) \equiv a_j$ $(2 \le j \le p)$ and $X_1 \equiv 1$, the model becomes a generalized partially linear model

$$g\{m(\mathbf{u},\, \mathbf{x})\} = a_1(\mathbf{u}) + \sum_{j=2}^{p} a_j\, x_j \tag{1.2}$$

studied, for example, by Chen (1988), Speckman (1988), Green and Silverman (1994), and Carroll, Fan, Gijbels and Wand (1997), among others. If we assume further that the function $a_1(\cdot)$ is also a constant, the model reduces to a familiar parametric generalized linear model (see McCullagh and Nelder 1989)

$$g\{m(\mathbf{u},\, \mathbf{x})\} = \sum_{j=1}^{p} a_j\, x_j. \tag{1.3}$$

In the least-squares setting, model (1.1) with the identity link was introduced by Cleveland, Grosse and Shyu (1992) and extended by Hastie and Tibshirani (1993) to various aspects. Furthermore, a two-step estimation procedure was proposed by Fan and Zhang (1997) to deal with the situations where coefficient functions admit different degrees of smoothness. An advantage of the model (1.1) is that via allowing coefficients $a_1(\cdot), \ldots, a_p(\cdot)$ to depend on $\mathbf{U}$, the modeling bias can be reduced significantly and the "curse of dimensionality" is avoided.

Generalized varying-coefficient models are a simple and useful extension of classical linear models. This extension admits simple interpretability. The models are particularly appealing in longitudinal studies where they allow one to explore the extent to which covariates affect responses changing over time. See Hoover *et al.* (1998), Brumback and Rice (1998) and Fan and Zhang (1998) for details on novel applications of the varying-coefficient models to longitudinal data. For nonlinear time series applications, see Chen and Tsay (1993) and Cai, Fan and Yao (1998) for statistical inferences based on functional-coefficient autoregressive models.

Estimation of coefficient functions in (1.1) is obtained by using local smoothing technique. By localizing data around the covariate $\mathbf{u}$, model (1.1) is approximately a generalized linear model and one can find its estimate by using a local maximum likelihood method. The local likelihood method relies on an iterative algorithm. In order to obtain estimated coefficient functions, we need to solve hundreds of local maximum likelihood problems. This can be expensive to compute, depending on the convergence criterion. Computational burden becomes even more severe when a cross-validation method is used to select a smoothing parameter. To attenuate this drawback, we propose a one-step local maximum likelihood estimator (MLE). Although the idea is not entirely new, our implementation is novel. Computational cost in a factor of tens can be saved and the resulting one-step estimator is demonstrated, both asymptotically and empirically, to be as efficient as the fully iterative MLE.

Associated with inferences on generalized varying-coefficient models are the standard errors of the estimated coefficient functions. Consistent estimates are derived. We further show that our estimated standard errors are indeed accurate enough for most of applications via empirical studies. Another important issue arises whether some of coefficient functions in model (1.1) are really varying, e.g., testing (1.3) against (1.1) or (1.2) versus (1.1), or whether some of covariates are statistically significant. A nonparametric maximum likelihood ratio test is proposed and its null distribution is estimated by using a conditional bootstrap method. Our simulation shows that the resulting testing procedure is indeed powerful and the bootstrap method gives the right null distribution.

The paper is organized as follows. In §2, generalized varying-coefficient models are introduced. §3 discusses estimation methods and inference tools. In particular, formulas for standard errors of estimated coefficient functions are derived, a maximum likelihood ratio test is proposed, and strategies are given for implementation of a one-step estimator. In §4, we study some finite sample properties of the one-step and local MLEs using two simulated examples. Furthermore, our methodology is illustrated through the aforementioned environmental dataset and a data set on survival probability of burn victims. §5 presents some asymptotic properties of the one-step and local MLEs. Finally, technical proofs are given in the Appendix.

2 Generalized varying-coefficient models

In generalized linear models, the conditional density of Y given covariate $(\mathbf{U}, \mathbf{X})$ belongs to the canonical exponential family:

$$f(y|\mathbf{u}, \mathbf{x}) = \exp\Big\{[\theta(\mathbf{u}, \mathbf{x})\, y - b\{\theta(\mathbf{u}, \mathbf{x})\}]/a(\phi) + c(y, \phi)\Big\} \tag{2.1}$$

for given functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. See McCullagh and Nelder (1989) and §5.4 of Fan and Gijbels (1996). Under model (2.1), one can easily show that the conditional mean and conditional variance are given respectively by $m(\mathbf{u}, \mathbf{x}) = E(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = b'\{\theta(\mathbf{u}, \mathbf{x})\}$, and $\mathrm{Var}(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = a(\phi)\, b''\{\theta(\mathbf{u}, \mathbf{x})\}$. Since our primary interest focuses on estimating the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{m(\mathbf{u}, \mathbf{x}), y\} = \theta(\mathbf{u}, \mathbf{x})\, y - b\{\theta(\mathbf{u}, \mathbf{x})\}.$$

Generalized varying-coefficient models extend the traditional generalized linear models by allowing coefficients to depend on a covariate. Namely, the linear predictor is

$$\eta(\mathbf{u}, \mathbf{x}) = g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p} a_j(\mathbf{u})\, x_j \tag{2.2}$$

for some given link function g. In practice, the canonical link $g(\cdot) = (b')^{-1}(\cdot)$ is frequently used.

Our goal is to estimate efficiently the coefficient functions $\{a_j(\cdot)\}$ by using a nonparametric method. Our methods are directly applicable to the situation where one can not specify fully the conditional likelihood function $\ell(v, y)$, but one can model the relationship between the mean function and variance function as $\mathrm{Var}(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = \sigma^2 V\{m(\mathbf{u}, \mathbf{x})\}$ for a given variance function $V(\cdot)$ and unknown σ. In this case, one needs only to replace the likelihood function $\ell(v, y)$ by the quasi-likelihood function $Q(\cdot, \cdot)$, defined by $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$.

3 Estimation

For simplicity, we consider only the case that $\mathbf{u}$ in (2.2) is one-dimensional. Extension to multivariate $\mathbf{u}$ involves no fundamentally new ideas. However, implementations with $\mathbf{u}$ more than two dimensions are not very useful due to the "curse of dimensionality".

3.1 Local MLE

We will use a local linear modeling scheme, though general local polynomial methods are also applicable. The local linear fittings have several nice properties such as high statistical efficiency (in an asymptotic minimax sense) and being design-adaptive (Fan 1993). Furthermore, they automatically correct edge effects (Ruppert and Wand 1994; and Fan and Gijbels 1996). Suppose that the second derivative of $a_j(\cdot)$ exists and is continuous. For each given point u_0, we approximate function $a_j(u)$ locally by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$ for u in a neighborhood of u_0.

Note that a_j and b_j depend on u_0. Based on a random sample $\{(U_i,\, \mathbf{X}_i,\, Y_i)\}_{i=1}^n$ from model (2.1), we use the following local likelihood method to estimate the coefficient functions

$$\ell(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[g^{-1} \left\{ \sum_{j=1}^{p} (a_j + b_j(U_i - u_0)) X_{ij} \right\},\, Y_i \right] K_h(U_i - u_0), \tag{3.1}$$

where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being a kernel function, $h = h_n > 0$ is a bandwidth, $\mathbf{a} = (a_1,\, \ldots,\, a_p)^T$ and $\mathbf{b} = (b_1,\, \ldots,\, b_p)^T$. Maximizing the local likelihood function $\ell(\mathbf{a},\, \mathbf{b})$ gives estimates $\widehat{\mathbf{a}}(u_0)$ and $\widehat{\mathbf{b}}(u_0)$. The components in $\widehat{\mathbf{a}}(u_0)$ give an estimate of $a_1(u_0),\, \ldots,\, a_p(u_0)$. For simplicity of notation, we will denote $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1,\, \ldots,\, a_p,\, b_1,\, \ldots,\, b_p)^T$ and write the local likelihood function (3.1) as $\ell(\boldsymbol{\beta})$. Likewise, the local MLE will be denoted by $\widehat{\boldsymbol{\beta}}_{\rm MLE}(u_0)$.

3.2 One-step local MLE

The local MLE can be costly to compute. This is particularly the case for the varying-coefficient models. In order to obtain functions $\{\widehat{a}_j(\cdot)\}$, one needs to maximize the local likelihood (3.1) for many (usually in an order of hundreds) distinct values of u_0. Each maximization requires an iterative algorithm. The computational cost of this iterative method depends also on the number of covariates p: the larger the more expensive. To reduce the computational cost, we propose to replace the iterative local MLE by an explicit non-iterative estimator. An excellent candidate is the one-step Newton-Raphson scheme, which has been frequently used in parametric models (see, for example, Bickel 1975; and Lehmann 1983). However, computational gain for the parametric models is not as significant as for our local likelihood estimation since the local likelihood method involves finding hundreds of parametric MLEs. Like in parametric models, it will be shown (see Theorem 2 below) that the one-step local MLE does not lose any statistical efficiency provided that the initial estimator is good enough.

We now describe our one-step iterative estimator, called one-step local MLE. Let $\ell'(\boldsymbol{\beta})$ and $\ell''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local likelihood $\ell(\boldsymbol{\beta})$. Given an initial estimator $\widehat{\boldsymbol{\beta}}_0(u_0) = \left(\widehat{\mathbf{a}}(u_0)^T,\, \widehat{\mathbf{b}}(u_0)^T\right)^T$, the Newton-Raphson algorithm is to find an updated estimator

$$\widehat{\boldsymbol{\beta}}_{\rm OS}(u_0) = \widehat{\boldsymbol{\beta}}_0(u_0) - \left\{ \ell'' \left(\widehat{\boldsymbol{\beta}}_0(u_0) \right) \right\}^{-1} \ell' \left(\widehat{\boldsymbol{\beta}}_0(u_0) \right). \tag{3.2}$$

This one-step estimator inherits clearly the computation expediency from least-squares local polynomial fitting.

In univariate generalized linear models, properties of the local one-step estimator were carefully studied by Fan and Chen (1999). In that setting, the least-squares estimate serves a natural candidate as an initial estimator. In the multivariate setting, however, it is not clear how an initial estimator can be constructed. Our implementation of the one-step local likelihood estimator is given in §3.5.

Note that $\ell''(\widehat{\boldsymbol{\beta}}_0(u_0))$ can be nearly singular for certain u_0, due to possible data sparsity in certain local regions. This is particularly the case when the bandwidth is small. A method to annihilate this drawback is the ridge regression. In the univariate setting, this idea was explored by Seifert and Gasser (1996) and Fan and Chen (1999). We will extend their ideas in §4.

3.3 Standard Errors

Standard errors are very useful for assessing sampling variability. It is frequently used in constructing pointwise confidence intervals. As shown in Theorem 2, the one-step estimator admits the same asymptotic variance as that of the local MLE. Therefore, the two estimators share the same estimated standard error.

Note that the local MLE (3.1) is really a weighted likelihood function of a corresponding parametric generalized linear model. Therefore, the covariate matrix can be estimated from the conventional technique. Let $q_j(s,\, y) = (\partial^j/\partial s^j)\, \ell\, \{g^{-1}(s),\, y\}$ and

$$\widehat{\Gamma}(u_0) = -\, \frac{1}{n}\, \sum_{i=1}^{n} q_2 \left[\sum_{j=1}^{p} \{\widehat{a}_j(u_0)\, X_{ij} + \widehat{b}_j(u_0)(U_i - u_0)\},\; Y_i \right] K_h(U_i - u_0) \begin{pmatrix} \mathbf{X}_i \\ \mathbf{V}_i \end{pmatrix}^{\otimes 2}, \tag{3.3}$$

where $\mathbf{V}_i = \mathbf{X}_i\, (U_i - u_0)/h$ and $A^{\otimes 2}$ denotes $A\, A^T$ for a matrix or vector A. Then, the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{i_0})$ can be estimated as

$$\widehat{\Sigma}^*(u_0) = \widehat{\Gamma}(u_0)^{-1}\, \widehat{\Lambda}(u_0)\, \widehat{\Gamma}(u_0)^{-1}, \tag{3.4}$$

where

$$\widehat{\Lambda}(u_0) = \frac{h}{n}\, \sum_{i=1}^{n} q_1^2 \left[\sum_{j=1}^{p} \{\widehat{a}_j(u_0)\, X_{ij} + \widehat{b}_j(u_0)(U_i - u_0)\},\; Y_i \right] K_h^2(U_i - u_0)\, \begin{pmatrix} \mathbf{X}_i \\ \mathbf{V}_i \end{pmatrix}^{\otimes 2}.$$

The estimated asymptotic variance of $\widehat{a}_j(u_0)$ is just the j^{th} diagonal element of $\widehat{\Sigma}^*(u_0)$. In our implementation, a ridge regression technique will be employed and hence the matrix $\widehat{\Gamma}(u_0)$ in (3.4) will be slightly modified to reflect this change.

The explicit formula for the asymptotic covariance matrix (see (5.3) below) provides an alternative estimate. The asymptotic covariance matrix is given by

$$\Sigma(u_0) = \mu_2\, \Gamma^{-1}(u_0)/f_{\mathrm{U}}(u_0), \tag{3.5}$$

where $\mu_k = \int u^k\, K(u)\, du$ for $k = 1$ and 2, $f_{\mathrm{U}}(\cdot)$ is the marginal density of U,

$$\Gamma(u) = E\left\{ \rho(U,\, \mathbf{X})\, \mathbf{X}\, \mathbf{X}^T \,|\, U = u \right\}, \tag{3.6}$$

and

$$\rho(u,\, \mathbf{x}) = [g_1\{m(u,\, \mathbf{x})\}]^2\, \mathrm{Var}\{Y | U = u,\, \mathbf{X} = \mathbf{x})\} \tag{3.7}$$

with $g_1(s) = g_0'(s)/g'(s)$ and $g_0(\cdot)$ being the canonical link. Note that $\rho(u,\, \mathbf{x}) = V\{m(u,\, \mathbf{x})\}$ for the canonical link function. Therefore, a direct estimate of $\Sigma(u_0)$ is $\widetilde{\Sigma}(u_0) = \mu_2\, \widehat{\Gamma}_S(u_0)^{-1}$, where $\widehat{\Gamma}_S(u_0)$ is the $p \times p$ upper corner submatrix of $\widehat{\Gamma}(u_0)$ given by (3.3).

3.4 Hypothesis testing

After fitting a generalized varying-coefficient model, one naturally asks whether the coefficient functions are really varying or whether any particular covariate is significant in the model. For simplicity of description, we only consider the first hypothesis testing problem

$$H_0 : a_1(u) \equiv a_1, \cdots, a_p(u) \equiv a_p, \tag{3.8}$$

though the technique also applies to other testing problems. A useful procedure is based on the nonparametric likelihood ratio test statistic

$$T = 2\{\ell(H_1) - \ell(H_0)\}, \tag{3.9}$$

where $\ell(H_0)$ and $\ell(H_1)$ are respectively the log-likelihood functions computed under the null hypothesis and the whole parametric space.

For parametric models, the likelihood ratio statistic follows asymptotically a χ^2-distribution with degrees of freedom $f - r$, where r and f are the number of parameters under the null and alternative hypotheses. For the nonparametric alternative, the effective number of parameters f tends to infinite. Thus, the test statistic will be asymptotically normal, *independent* of the values $a_1, \cdots, a_p$. This in turn suggests that we can use the following conditional bootstrap to construct the null distribution of T. Let $\widehat{a}_1, \cdots, \widehat{a}_p$ be the MLE under the null hypothesis. Given the covariates $(U_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, generate a bootstrap sample Y_i^* from model (2.1) with

$$\widehat{\eta}(U_i, \mathbf{X}_i) = \sum_{j=1}^{p} \widehat{a}_j X_{ij}$$

and compute the test statistic T^* in (3.9). Use the distribution of T^* as an approximation to the distribution of T. This method is valid since the asymptotic null distribution does not depend on the values of $a_1, \cdots, a_p$. The statement will be verified in §4.

Note that the above conditional bootstrap method applies readily to the Poisson and Bernoulli distributions, since in these cases (2.1) does not involve with any dispersion parameters. It is really a simulation approximation to the conditional distribution of T given observed covariates under the particular null hypothesis: $H_0 : a_j(u) = \widehat{a}_j$ $(j = 1, \cdots, p)$. As pointed above, this approximation is valid under both H_0 and H_1 as the null distribution does not asymptotically depend on the values of $a_j (j = 1, \cdots, p)$. In the case where model (1.1) involves a dispersion parameter (e.g., the Gaussian model), the dispersion parameter should be estimated based on the residuals from the *alternative* hypothesis.

For testing the hypothesis such as $a_p(\cdot) = 0$. The above conditional bootstrap idea continues to apply. In this case, the data should be generated from the mean function

$$g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p-1} \widehat{a}_j(\mathbf{u}) x_j,$$

where $\widehat{a}_j(\cdot)$ is an estimate under the alternative hypothesis.

3.5 Implementation of one-step local MLE

Suppose that we wish to evaluate the functions $\widehat{\mathbf{a}}(\cdot)$ at grid points $u_j,\ j = 1, \ldots, n_{\text{grid}}$. Our idea of finding initial estimators is as follows. Take a point u_{i_0}, usually the center of the grid points. Compute the local MLE $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{i_0})$. Use this estimate as the initial estimate for the point u_{i_0+1} and apply (3.2) to obtain $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0+1})$. Now, use $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0+1})$ as the initial estimate at the point u_{i_0+2} and apply (3.2) to obtain $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0+2})$ and so on. Likewise, we can compute $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0-1}),\ \widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0-2})$, etc. In this way, we obtain our estimates at all grid points.

There are a couple of possible variations to the above technique. The first one is to calculate a fresh local MLE as a new initial value after iterating along the grid points for a while. For example, if we wish to evaluate the functions at 200 grid points and are willing to compute the local maximum likelihood at five distinct points. A sensible placement of these points is $u_{20},\ u_{60},\ u_{100},\ u_{140}$ and u_{180}. Use for example $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{60})$ along with the idea in the last paragraph to compute $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_i)$ for $i = 40, \ldots, 79$. In our implementation, this modified technique is used.

Another useful modification is to use a two-step method. We use the scenarios given in the last paragraph as an illustration. After obtaining $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{60})$, say, we apply (3.2) to obtain $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{61})$. Regarding $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{61})$ as an initial value, we use (3.2) to obtain a "two-step" estimator $\widehat{\boldsymbol{\beta}}_{\text{TS}}(u_{61})$. Now, use $\widehat{\boldsymbol{\beta}}_{\text{TS}}(u_{61})$ as an initial value for the grid point u_{62} and iterate (3.2) twice to obtain $\widehat{\boldsymbol{\beta}}_{\text{TS}}(u_{62})$ and so on. This implementation requires approximately twice as much effort to compute the estimates as the one-step method. However, our empirical studies show that there are no significant differences between the two procedures. See §4 for details.

The theoretical basis for the above "one-step" and the "two-step" procedures is as follows. When the grid points are sufficient fine, $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{i_0})$ will be very close to $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{i_0+1})$. Indeed, when the grid span is of order $O\left\{h_n^2 + (n\,h_n)^{-1/2}\right\}$ which usually is true for most applications, $\widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_{i_0})$ satisfies the condition given in Theorem 2. Therefore, $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0+1})$ is as efficient as the fully-iterative local MLE at the point u_{i_0+1}. Using the same reasoning, $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{i_0+2})$ is as efficient as the local MLE at the point $u = u_{i_0+2}$ and so on. The same arguments are still applicable for the two-step estimator. A refresh start is needed because of stochastic error accumulation as iterations along grid points march on.

4 Simulations and applications

In this section, we first discuss how to implement the one-step procedure for two important models, the Bernoulli and the Poisson models. We then illustrate the performance of the proposed one-step method and compare it with the two-step estimator and the fully-iterative local MLE. The performance of estimator $\widehat{a}_j(\cdot)$ is assessed via the square-Root of Average Square Errors (RASE):

$$\text{RASE}_j^2 = n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \{\widehat{a}_j(u_k) - a_j(u_k)\}^2, \qquad j = 1, \cdots, p, \tag{4.1}$$

where $\{u_j, j = 1, \ldots, n_{grid}\}$ are the grid points at which the functions $\{a_j(\cdot)\}$ are estimated. Similarly the performance of the joint estimator $\widehat{\mathbf{a}}(u)$ is evaluated by

$$\mathrm{RASE}^2 = \sum_{j=1}^{p} \mathrm{RASE}_j^2. \tag{4.2}$$

In the following two simulated examples, the covariates X_1 and X_2 are standard normal random variables with correlation coefficient $2^{-1/2}$ and U is uniformly distributed over $[0, 1]$, independent of $(X_1,\ X_2)$. Three bandwidths will be employed, which represent approximately the situations of undersmooth, about right amount of smooth and oversmooth. For this wide range of bandwidths, we compare the performances among the one-step, the two-step and the fully iterative local MLE methods. The Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and $n_{\mathrm{grid}} = 200$ are used.

4.1 Logistic Regression

For a Bernoulli distribution with a logit link, the local likelihood $\ell(\mathbf{a},\ \mathbf{b})$ in (3.1) now becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i\sum_{j=1}^{p}\{a_j + b_j(U_i - u_0)\}X_{ij} - \log\left\{1 + \exp\left(\sum_{j=1}^{p}(a_j + b_j(U_i - u_0))X_{ij}\right)\right\}\right] K_h(U_i - u_0).$$

and the one-step estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} = \widehat{\boldsymbol{\beta}}_0 + \begin{pmatrix} \mathbf{H}_{n,0}, & \mathbf{H}_{n,1} \\ \mathbf{H}_{n,1}, & \mathbf{H}_{n,2} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{v}_{n,0} \\ \mathbf{v}_{n,1} \end{pmatrix}, \tag{4.3}$$

where

$$\begin{aligned} \mathbf{H}_{n,j} &= \textstyle\sum_{i=1}^{n} K_h(U_i - u_0)\widehat{p}_{i0}(1 - \widehat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i\mathbf{X}_i^T, \qquad j = 0,\ 1,\ 2, \\ \mathbf{v}_{n,j} &= \textstyle\sum_{i=1}^{n} K_h(U_i - u_0)(Y_i - \widehat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i, \qquad j = 0, 1 \end{aligned}$$

with $\widehat{p}_{i0}$ satisfying

$$\mathrm{logit}\,(\widehat{p}_{i0}) = \sum_{j=1}^{p}\left\{\widehat{a}_{j,0} + \widehat{b}_{j,0}(U_i - u_0)\right\} X_{ij}.$$

The two-step estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{TS}}$ is obtained by iterating twice the equation (4.3) and the local MLE is simply iterated using equation (4.3) until convergence.

In practical implementations, the matrix in (4.3) can be singular or nearly singular when the local data are sparse. To attenuate the difficulty, one may follow the idea of ridge regression (Seifert and Gasser 1996; and Fan and Chen 1999). Then an issue arises on how to choose the ridge parameters. Note that the k-th diagonal element of $\mathbf{H}_{n,j}$ ($j = 0$ and 2) is approximately of order

$$E\left(X_k^2 \,|\, U = u_0\right)\, \widehat{p}_0(1 - \widehat{p}_0)\, h^{j-1} \int u^j K(u)\, du\ N \qquad \text{with} \qquad \widehat{p}_0 = \frac{\exp(\widehat{\mathbf{a}}_0^T\overline{\mathbf{X}})}{1 + \exp(\widehat{\mathbf{a}}_0^T\overline{\mathbf{X}})}, \tag{4.4}$$

where $N = n\,h\,f_{\mathrm{U}}(u_0)$ and $\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$. The parameter N can be intuitively understood as the effective number of local data points. This motivates us to use the ridge parameter

$$r_{j,k} = \left(\frac{1}{n}\sum_{i=1}^{n} X_{ik}^2\right)\, \widehat{p}_0(1 - \widehat{p}_0)h^{j-1} \int u^j K(u)\, du$$

for the k-th diagonal element of $\mathbf{H}_{n,j}$. Using such a ridge parameter will not alter the asymptotic behavior and will prevent the matrix from nearly singular when N is small.

Example 1. Take $\mathbf{X} = (1,\ X_1,\ X_2)^T$ and the coefficient functions in (1.1) are given by

$$a_0(u) = \exp(2u - 1), \quad a_1(u) = 8u(1-u), \quad \text{and} \quad a_2(u) = 2\ \sin^2(2\pi u). \tag{4.5}$$

Figure 1(a) depicts the marginal distributions for the ratios of the overall RASE defined in (4.2), using three bandwidths $h = 0.1$, 0.2 and 0.4. The ratios based on individual RASE defined in (4.1) were also computed and the results are not reported here since they are similar to Figure 1(a). It is evident that the performance of the one-step, the two-step and the fully iterative estimators are comparable for a wide range of bandwidths. As expected, the performance of the two-step estimator is closer to that of the local MLE. Figures 1(b)–(d) give a typical estimate of the coefficient functions. Table 1 summarizes the simulation results with μ and σ denoting the mean and standard deviation

Table 1. Bivariate summary of simulation results for logistic regression model

		MLE		One-step			Two-step		
n	h	μ	σ	μ	σ	ρ_*	μ	σ	ρ_*
400	0.10	2.2278	2.0874	1.8537	0.9759	0.8656	2.1244	1.5315	0.8274
	0.20	1.0669	0.4491	1.0576	0.4378	0.9991	1.0669	0.4491	1.0000
	0.40	0.9454	0.1600	0.9447	0.1593	1.0000	0.9454	0.1600	1.0000
800	0.075	1.2451	0.6639	1.1644	0.3767	0.8342	1.2256	0.5301	0.9656
	0.15	0.7280	0.2573	0.7234	0.2459	0.9993	0.7280	0.2573	1.0000
	0.30	0.7433	0.1009	0.7429	0.1005	1.0000	0.7433	0.1009	1.0000

of the RASE in 400 simulations. Here, ρ_* indicates the correlation coefficient between the RASE of the MLE and the RASE of the one-step (or two-step) method. Note that the correlation coefficients are close to one which indicates that the one-step and two-step methods follow closely the MLE. Note also that the larger the bandwidths, the larger the correlation coefficients. This is due to the fact that a larger bandwidth implies more local data points, which makes the asymptotic theory more relevant. As expected, the correlation coefficients for the two-step method are larger than those of the one-step method, since the former is closer to the MLE.

We now test the accuracy of our standard error formula (3.4). The standard deviation, denoted by SD in Table 2, of 400 estimated $\widehat{a}_j(u_0)$, based on 400 simulations, can be regarded as the true standard errors. The average and the standard deviation of 400 estimated standard errors, denoted by SD_a and SD_{std}, summarize the overall performance of the standard error formula (3.4). Table 2 presents the results at the points $u_0 = 0.25$, 0.50 and 0.75. It suggests that our standard error formula somewhat underestimates the true standard deviation, though the difference is within two standard deviations of the Monte Carol errors. The bias becomes smaller as the number of local data points $n\,h_n$ goes up (see the last two situations). This is consistent with our asymptotic theory.

Next, we conduct a simulation study to verify the statements that the asymptotic null distribution of the test statistic T defined in (3.9) does not depend on the values of $\{a_j\}$ under H_0 (see (3.8)) and that the limiting conditional null distributions are independent of the covariate values.

Table 2. Standard deviations of estimators for logistic regression model

			$\widehat{a}_0(u)$		$\widehat{a}_1(u)$		$\widehat{a}_2(u)$	
n	h	u	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})
400	0.2	0.25	0.3185	0.2673 (0.0470)	0.4890	0.4069 (0.0776)	0.5082	0.3986 (0.0893)
		0.50	0.3410	0.2782 (0.0451)	0.5413	0.4330 (0.0809)	0.4135	0.3568 (0.0591)
		0.75	0.4315	0.3542 (0.0776)	0.5372	0.4542 (0.0996)	0.5809	0.4431 (0.0969)
400	0.3	0.25	0.2294	0.2051 (0.0231)	0.3424	0.3201 (0.0447)	0.3317	0.2956 (0.0403)
		0.50	0.2570	0.2315 (0.0315)	0.3931	0.3538 (0.0527)	0.3490	0.3122 (0.0431)
		0.75	0.2850	0.2686 (0.0423)	0.3929	0.3581 (0.0557)	0.3788	0.3328 (0.0500)
800	0.15	0.25	0.2418	0.2214 (0.0214)	0.3638	0.3460 (0.0501)	0.3804	0.3486 (0.0532)
		0.50	0.2249	0.2196 (0.0233)	0.4040	0.3569 (0.0512)	0.3124	0.2812 (0.0356)
		0.75	0.3146	0.2928 (0.0478)	0.4209	0.3804 (0.0667)	0.3987	0.3781 (0.0631)

To this end, we compute the unconditional null distribution of T with $n = 400$, via 1000 Monte Carlo simulations, for 5 different sets of values of $\{a_j\}$. These sets of parameters are quite far apart. The resulting 5 densities are depicted in Figure 1(e) (thick curves). They are nearly the same, which suggest that the asymptotic null distribution does not depend on the values of $\{a_j\}$. To validate our conditional bootstrap method, five typical data sets were selected from our previous 400 simulations. The estimated conditional bootstrap null distributions, based on 1000 bootstrap samples, are plotted as thin curves in Figure 1(e). Six empirical percentiles for five different sets of values of $\{a_j\}$ and covariates are listed in Table 3. Both Figure 1(e) and Table 3 shows that they

Table 3. Six empirical percentiles for logistic model

10	25	50	75	90	95
Conditional bootstrap					
7.9579	10.7189	14.2569	18.2625	22.2566	24.9903
8.2450	11.0170	14.6601	18.4897	22.4177	25.5829
8.0004	10.9871	14.2667	18.0413	22.5517	25.1661
8.7738	11.4311	14.8061	18.5209	22.7029	25.3781
8.7906	11.4672	14.9130	18.6168	22.3256	24.7104
Unconditional bootstrap					
7.6381	10.7167	14.5487	18.6276	22.2205	24.4597
7.3478	10.1290	13.9934	17.9622	21.8270	24.4429
7.7238	11.3849	14.6151	18.4796	22.5899	24.7270
8.8042	11.3762	14.8076	18.7571	22.0560	25.1550
8.7865	11.3472	14.5975	18.5198	23.1476	25.8297

are very close to the true null distribution. This demonstrates that our bootstrap method gives the correct null distribution even when the data were generated from an alternative model (4.5).

To examine the power of the proposed test, we consider the following null hypothesis

$$H_0 : a_j(u) = \theta_j, \quad j = 0, 1, 2,$$

namely a generalized linear model, versus the alternative

$$H_1 : a_j(u) \neq \theta_j, \text{ for at least one } j.$$

The power functions are evaluated under a sequence of the alternative models indexed by β

$$H_1 : a_j(u) = a_{j0} + \beta(a_j^0(u) - a_{j0}), \quad j = 0,\ 1,\ 2 \quad (0 \leq \beta \leq 0.8),$$

where $\{a_j^0(u)\}$ are given in (4.5) and $a_{j0} = E\{a_j(U)\}$. Figure 1(f) depicts the five power functions based on 1000 simulations for the sample size $n = 400$ at five different significant levels: 0.5, 0.25, 0.10, 0.05, and 0.01. When $\beta = 0$, the special alternative collapses into the null hypothesis. The powers at $\beta = 0$ for the above 5 significant levels are respectively 0.532, 0.281, 0.101, 0.047 and 0.012. This shows that the conditional bootstrap method gives the right levels of test. The power functions increase rapidly as β increases. This in turn shows that the test proposed in §3.4 is indeed powerful.

4.2 Poisson regression

For a Poisson model with the canonical link, the log-likelihood function is given by

$$\ell(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} K_h(U_i - u_0)\left[Y_i\left\{\sum_{j=1}^{p}(a_j + b_j(U_i - u_0))X_{ij}\right\} - \exp\left\{\sum_{j=1}^{p}(a_j + b_j(U_i - u_0))X_{ij}\right\}\right].$$

By straightforward calculation, the one-step estimator is given similarly to (4.3) but now

$$\mathbf{H}_{n,j} = \sum_{i=1}^{n} K_h(U_i - u_0)\widehat{\lambda}_{i0}(U_i - u_0)^j \mathbf{X}_i\mathbf{X}_i^T, \quad j = 0,\ 1,\ 2,$$

and

$$\mathbf{v}_{n,j} = \sum_{i=1}^{n} K_h(U_i - u_0)(Y_i - \widehat{\lambda}_{i0})(U_i - u_0)^j \mathbf{X}_i,\ \ j = 0, 1,$$

where $\widehat{\lambda}_{i0} = \exp\left[\sum_{j=1}^{p}\{\widehat{a}_{j0} + \widehat{b}_{j0}(U_i - u_0)\}X_{ij}\right]$. Using the same arguments as in the previous section, the ridge parameters

$$r_{j,k} = \left(\frac{1}{n}\sum_{i=1}^{n} X_{ik}^2\right)\widehat{\lambda}_0\, h^{j-1}\int u^j K(u)\, du \qquad \text{with} \qquad \widehat{\lambda}_0 = \exp\left(\widehat{\mathbf{a}}_0^T\overline{\mathbf{X}}\right) \tag{4.6}$$

are employed against the possible singularity of matrix $\mathbf{H}_{n,j}$ ($j = 0$ and 2) in (4.3).

Example 2. The conditional distribution of Y given covariates U, X_1 and X_2 is taken to be Poisson with the following linear predictor

$$\eta(u,\ \mathbf{x}) = 5.5 + 0.1\{a_0(u) + a_1(u)\, x_1 + a_2(u)\, x_2\},$$

where the coefficient functions $a_0(u)$, $a_1(u)$ and $a_2(u)$ are the same as those in Example 1. The coefficients 5.5 and 0.1 are chosen so that the range of simulated data is close to that of the environmental data in §4.3.

Figure 2 and Table 4 summarize the result for $n = 200$. It shows again that the one-step, two-step and the iterative local MLE have comparable performance. A typical estimated function with bandwidth $h = 0.15$ is presented in Figures 2(b)–(d). Because of different noise-to-signal ratios, the functions here are indeed estimated better than those given in Example 1. Similar to Example 1, we summarize the performance of our estimated standard error formula (3.4) in Table 5. Clearly, our estimated standard errors are very close to the true ones.

Table 4. Bivariate summary of simulation output for Poisson regression model

		MLE		One-step			Two-step		
n	h	μ	σ	μ	σ	ρ_*	μ	σ	ρ_*
200	0.075	0.3632	0.0692	0.3468	0.0562	0.8691	0.3632	0.0692	1.0000
	0.15	0.3220	0.0510	0.3202	0.0504	0.9925	0.3220	0.0510	1.0000
	0.30	0.5852	0.0425	0.5835	0.0426	0.9990	0.5852	0.0425	1.0000
400	0.075	0.2309	0.0352	0.2279	0.0347	0.9866	0.2309	0.0352	1.0000
	0.15	0.2581	0.0325	0.2571	0.0322	0.9942	0.2581	0.0325	1.0000
	0.30	0.5603	0.0292	0.5581	0.0293	0.9988	0.5603	0.0292	1.0000

Table 5. Standard deviations of estimators for Poisson regression model

			$\widehat{a}_0(u)$		$\widehat{a}_1(u)$		$\widehat{a}_2(u)$	
n	h	u	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})
200	0.15	0.25	0.0105	0.0092 (0.0013)	0.0148	0.0118 (0.0024)	0.0156	0.0126 (0.0026)
		0.50	0.0094	0.0088 (0.0011)	0.0148	0.0112 (0.0022)	0.0150	0.0118 (0.0024)
		0.75	0.0100	0.0088 (0.0011)	0.0142	0.0112 (0.0023)	0.0151	0.0119 (0.0023)
400	0.075	0.25	0.0094	0.0085 (0.0012)	0.0130	0.0106 (0.0021)	0.0136	0.0107 (0.0022)
		0.50	0.0093	0.0083 (0.0011)	0.0127	0.0104 (0.0022)	0.0130	0.0105 (0.0021)
		0.75	0.0090	0.0081 (0.0011)	0.0137	0.0101 (0.0022)	0.0133	0.0102 (0.0022)

Similar to Example 1, the procedure of testing hypothesis is applied to this example. Both unconditional and conditional estimated densities of T are displayed in Figure 2(e). Six empirical percentiles are listed in Table 6. The corresponding power functions are presented in Figure 2(f). The same conclusions as those in Example 1 can be drawn for the Poisson regression models. In particular, the test has the correct levels of significance. See the power functions in Figure 2(e) at $\beta = 0$.

4.3 Real-data examples

Example 3. We in this example illustrate our proposed procedure via an application to the environmental data set mentioned in the introduction. Of interest is to study the association between levels of pollutants and number of total hospital admissions for circulationary and respirationary

Table 6. Six empirical percentiles for Poisson model

10	25	50	75	90	95
Conditional bootstrap					
12.1646	15.1401	18.6981	22.6260	26.1432	28.8494
11.7506	14.5010	18.0994	22.3809	26.1237	29.4936
11.7946	14.7005	18.3495	22.2918	26.0064	29.2165
11.4662	14.6917	18.2475	22.4623	27.0587	29.6887
11.9894	14.7869	18.5571	22.3593	26.7014	29.7923
Unconditional bootstrap					
11.9492	14.7920	18.5509	22.3383	26.7474	28.8094
11.1599	14.7156	18.7054	22.2915	26.6170	28.9831
11.4378	14.8132	18.4080	22.3890	26.5858	29.4816
11.8238	14.6817	18.5090	22.7050	26.4776	29.3814
11.8365	14.9721	18.7674	22.9402	26.5929	28.9815

problems on every Friday from January 1, 1994 to December 31, 1995 and to examine the extent to which the association varies over time. The covariates are taken as the levels of pollutants Sulfur Dioxide X_2 (in $\mu g/m^3$), Nitrogen Dioxide X_3 (in $\mu g/m^3$) and dust X_4 (in $\mu g/m^3$). It is reasonable to use the Poisson regression model with the mean $\lambda(t,\, \mathbf{x})$ given by

$$\log\{\lambda(t,\, \mathbf{x})\} = a_1(t) + a_2(t)x_2 + a_3(t)x_3 + a_4(t)x_4. \tag{4.7}$$

Both the one-step and local likelihood methods were employed to estimate the coefficient functions $a_j(\cdot)$ and the results are similar.

A multifold cross-validation method was used to select a bandwidth. We partitioned the data into 20 groups — the j^{th} group consisting of data points with indices

$$d_j = \{20k + j, k = 1, 2, \cdots\}, \quad j = 0, \cdots, 19.$$

For each given j, the j-th group of data were deleted and the model (4.7) was fitted for the remaining data. Then the deviance (see, e.g., page 34 of McCullagh and Nelder 1989) or the sum of squares of Pearson's residuals were computed. This leads to two cross-validation criteria:

$$CV_1(h) = \sum_{j=0}^{19} \sum_{i \in d_j} 2\left[y_i \log\{y_i/\widehat{y}_{-d_j}(U_i,\, \mathbf{X}_i)\} - \{y_i - \widehat{y}_{-d_j}(U_i,\, \mathbf{X}_i)\}\right],$$

and

$$CV_2(h) = \sum_{j=0}^{19} \sum_{i \in d_j} \left\{ \frac{y_i - \widehat{y}_{-d_j}(U_i,\, \mathbf{X}_i)}{\sqrt{\widehat{y}_{-d_j}(U_i,\, \mathbf{X}_i)}} \right\}^2,$$

where $\widehat{y}_{-d_j}(U_i,\, \mathbf{X}_i)$ is a fitted value with the data in d_j deleted. Figure 3(b) depicts the cross-validation functions $CV_1(h)$ and $CV_2(h)$ and results in the optimal bandwidth $h = 0.1440 \times 105$. The estimated coefficient functions are summarized in Figure 4. They describe the extent to which the association between the pollutants and the number of hospital admissions vary over time. The figure shows clearly that the coefficient functions vary with time. The two dashed curves are the estimated function plus/minus twice of the estimated standard errors. They give us an idea of the pointwise confidence intervals with bias ignored.

A question arises whether or not the data are highly correlated. To check for the serial correlation, Pearson's residuals are computed. The time series plot of the residuals is given in Figure 5(a) and the plot of the corresponding autocorrelation coefficients against time lag is presented in Figure 5(b). There is no pattern in Figure 5(a), which, together with Figure 5(b), concludes that there is no evidence that the data are serially correlated.

We now apply the procedure proposed in §3.4 to testing whether the coefficients are really time varying. The MLE under the null hypothesis is (5.4499, -0.0025, 0.0015, -0.0005) with an estimated standard deviation (0.0195, 0.0006, 0.0006, 0.0005). The test statistic (3.9) is $T = 389.41$, which suggests that varying-coefficient model is a much better fit. Based on 1000 bootstrap replications, the distribution of T is estimated (see Figure 6). The sample mean and sample variance of T^* are 26.64 and 48.40, respectively, which suggests that the underlying distribution

may be approximated by a χ^2 distribution with degrees of freedom 27 (see Figure 6). The p-value is zero, which strongly rejects the null hypothesis.

The parametric Poisson model suggests that the dust level (X_4) is not statistically significant. We would have concluded that X_4 can be deleted from the parametric fit should the parametric Poisson model be used. To examine if the variable X_4 is significant in the generalized varying-coefficient model, we apply the idea in §3.4 to testing the hypothesis: the function $a_4(\cdot)$ is zero. The maximum likelihood ratio test statistic is $T = 20.1847$. Based on 1000 bootstrap samples, the p-value is 0.321 (the sample mean and variance of T^* are 17.7352 and 37.1976, respectively). Therefore, the variable X_4 should be dropped from the generalized varying-coefficient model. After deleting the variable dust level (X_4), the MLE for the parametric Poisson model is (5.4523, −0.0025, 0.0010) with an estimated standard deviation (0.0193, 0.0006, 0.0004), which implies that both covariates Sulfur Dioxide (X_2) and Nitrogen Dioxide (X_3) are statistically significant. Finally, we apply the same procedure as above to test whether X_3 is statistically significant in the generalized varying-coefficient model. That is to test $H_0 : \log\{\lambda(t, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2$ against $H_1 : \log\{\lambda(t, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2 + a_3(t)\, x_3$. As a result, the maximum likelihood ratio test statistic is $T = 39.7473$ and the p-value is 0.039 (the sample mean and variance of T^* are 27.5071 and 39.5808, respectively), based on 1000 bootstrap samples. Therefore, the variable Nitrogen Dioxide (X_3) is significant at level 0.05. The conclusion is consistent with the parametric analysis.

Example 4. Now we apply the methodology proposed in this paper to analyze the data set: *Burns data*, collected by General Hospital Burn Center at the University of Southern California. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, and covariates X_1=*age*, X_2=*sex*, $X_3 = \log(\text{burn area}+1)$ and binary variable X_4=*Oxygen* (0 normal, 1 abnormal) are considered. We are interested in studying how burn areas and the other variables affect survival probabilities for victims at different age groups. This leads to naturally following varying-coefficient model:

$$logit\{p(x_1, x_2, x_3, x_4)\} = a_1(x_1) + a_2(x_1)x_2 + a_3(x_1)x_3 + a_4(x_1)x_4. \tag{4.8}$$

Figure 7 presents the estimated coefficients for model (4.8) via the one-step approach with bandwidth $h = 65.7882$, selected by a cross-validation method.

A natural question arises whether the coefficients in (4.8) are really varying. To see this, we consider the parametric logistic regression model

$$logit\{p(x_1, x_2, x_3, x_4)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{4.9}$$

as the null model. The MLE of $(\beta_0, \cdots, \beta_4)$ and its standard deviation for model (4.9) are $(23.2213, -6.1485, -0.4661, -2.4496, -0.9683)$ and $(1.9180, 0.6647, 0.2825, 0.2206, 0.2900)$, respectively. The test statistic T proposed in §3.4 is 54.9601 with p-value 0.000, based on 1000 bootstrap samples (the sample mean and variance of T^* are 5.9756 and 10.7098, respectively). This implies that the varying-coefficient logistic regression model fits the data much better than the parametric fit. It also allows us to examine the extent to which the regression coefficients vary over different ages.

From Figure 7, it can be observed that both functions $a_2(\cdot)$ and $a_4(\cdot)$ are nearly constant. This leads us to testing hypothesis H_0 : both $a_2(\cdot)$ and $a_4(\cdot)$ are constant under model (4.8). The corresponding test statistic T is 3.2683 with p-value 0.7050, based on 1000 bootstrap samples. This in turn suggests that the coefficient functions $a_2(\cdot)$ and $a_4(\cdot)$ are independent of age and indicates that there are no gender differences for different age groups.

Finally, we examine whether both covariates *sex* and *Oxygen* are statistically significant in model (4.8). The likelihood ratio test for this problem is $T = 11.2727$ with p-value 0.0860, based on 1000 bootstrap samples (the sample mean and variance of T^* are 5.2867 and 9.7630, respectively). Both covariates *sex* and *Oxygen* are not significant at level 0.05. This is intuitively expected: gender and oxygen do not play a significant role in determining the survival probability of a victim.

5 Asymptotic theory

In this section, we derive the asymptotic distributions of the local MLE $\widehat{\boldsymbol{\beta}}_{\rm MLE}$ and the one-step estimator $\widehat{\boldsymbol{\beta}}_{\rm OS}$. We demonstrate that the one-step estimator performs as well as the local MLE as long as the initial estimator $\widehat{\boldsymbol{\beta}}_0$ is reasonably accurate (see (5.4) below). In other words, the one-step estimator reduces computational cost of the local MLE without downgrading its asymptotic performance.

Denote by $\nu_k = \int u^k\, K^2(u)\, du$ for $k = 0$, 1, and 2. Let $\mathbf{H} = \mathrm{diag}(1,\, h) \otimes \mathbf{I}_p$ with $\otimes$ denoting the Kronecker product. Now we state our theorems here but their proofs are relegated in the Appendix. Also the conditions for the theorems are listed in the Appendix.

Theorem 1. Suppose that Conditions (1) – (7) in the Appendix hold and that $h = h_n \to 0$ and $n\,h \to \infty$ as $n \to \infty$. Then

$$\sqrt{n\,h}\left[\mathbf{H}\left\{\widehat{\boldsymbol{\beta}}_{\rm MLE}(u_0) - \boldsymbol{\beta}(u_0)\right\} - \frac{h^2}{2(\mu_2 - \mu_1^2)}\begin{pmatrix}(\mu_2^2 - \mu_1\,\mu_3)\,\mathbf{a}''(u_0)\\ (\mu_3 - \mu_1\,\mu_2)\,\mathbf{a}''(u_0)\end{pmatrix} + o_p(h^2)\right]$$
$$\xrightarrow{\mathcal{D}} \quad N\left(0,\ \Delta^{-1}\,\Lambda\,\Delta^{-1}\right), \tag{5.1}$$

where with $\Gamma(u_0)$ given by (3.7),

$$\Delta = f_{\rm U}(u_0)\begin{pmatrix}1 & \mu_1\\ \mu_1 & \mu_2\end{pmatrix} \otimes \Gamma(u_0) \qquad \text{and} \qquad \Lambda = f_{\rm U}(u_0)\begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix} \otimes \Gamma(u_0). \tag{5.2}$$

Furthermore, if $K(\cdot)$ is symmetric,

$$\sqrt{n\,h}\ \left[\widehat{\mathbf{a}}_{\rm MLE}(u_0) - \mathbf{a}(u_0) - \frac{h^2\,\mu_2}{2}\,\mathbf{a}''(u_0) + o_p(h^2)\right] \quad \xrightarrow{\mathcal{D}} \quad N\left(0,\ \Sigma(u_0)\right), \tag{5.3}$$

where $\Sigma(u_0)$ is defined in (3.5).

Theorem 2. Under the assumptions in Theorem 1, then $\widehat{\boldsymbol{\beta}}_{\rm OS}$ has the same asymptotic distribution as $\widehat{\boldsymbol{\beta}}_{\rm MLE}$, provided that the initial estimator $\widehat{\boldsymbol{\beta}}_0$ satisfies

$$\mathbf{H}\,\left(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) = O_p\left\{h^2 + (n\,h)^{-1/2}\right\}. \tag{5.4}$$

In other words, (5.1) and (5.3) hold true for $\widehat{\boldsymbol{\beta}}_{\rm OS}$.

References

Bickel, P.J. (1975), "One-step Huber estimates in linear models," *Journal of the American Statistical Association*, **70**, 428-433.

Brumback, B. and Rice, J. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves," *Journal of the American Statistical Association*, **93**, 961–976.

Cai, Z., Fan, J. and Yao, Q. (1998), "Functional-coefficient regression models for nonlinear time series," revised for *Journal of the American Statistical Association*.

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

Chen, H. (1988), "Convergence rates for parametric components in a partly linear model," *The Annals of Statistics*, **16**, 136-146.

Chen, R. and Tsay, R.S. (1993), "Functional-coefficient autoregressive models," *Journal of the American Statistical Association*, **88**, 298-308.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992), "Local regression models," in *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309–376. Pacific Grove, California: Wadsworth & Brooks.

Fan, J. (1993), "Local linear regression smoothers and their minimax," *The Annals of Statistics*, **21**, 196–216.

Fan, J. and Chen, J. (1999), "One-step local quasi-likelihood estimation," *Journal of the Royal Statistics Society, Series B*, to appear.

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

Fan, J. and Zhang, J. (1998), "Functional linear models for longitudinal data," Institute of Statistics Mimeo Series, University of North Carolina.

Fan, J. and Zhang, W. (1997), "Statistical estimation in varying-coefficient models," revised for *The Annals of Statistics*.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models: A Robust Penalty Approach*. London: Chapman and Hall.

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

Hastie, T.J. and Tibshirani, R. J. (1993), "Varying-coefficient models (with discussion)," *Journal of the Royal Statistical Society, Series B*, **55**, 757-796.

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

Lehmann, E.L. (1983), *Theory of Point Estimation*. Pacific Grove, California: Wadsworth & Brooks/Cole.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

Ruppert, D. and Wand, M.P. (1994), "Multivariate weighted least squares regression," *The Annals of Statistics*, **22**, 1346–1370.

Seifert, B. and Gasser, Th. (1996), "Finite-sample variance of local polynomial: Analysis and solutions," *Journal of the American Statistical Association*, **91**, 267-275.

Speckman, P. (1988), "Kernel smoothing in partial linear models," *Journal of the Royal Statistical Society, Series B*, **50**, 413-436.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*. London: Chapman and Hall.

Appendix: Proofs

We first impose some regularity conditions. Note that $q_k(\cdot,\,\cdot)$ is linear in y for fixed s such that

$$q_1[g\{m(u,\,\mathbf{x})\},\,m(u,\,\mathbf{x})] = 0 \qquad \text{and} \qquad q_2[g\{m(u,\,\mathbf{x})\},\,m(u,\,\mathbf{x})] = -\rho(u,\,\mathbf{x}), \tag{A.1}$$

where $\rho(u,\,\mathbf{x})$ is defined in (3.7).

Conditions:

(1) The function $q_2(s,\,y) < 0$ for $s \in \Re$ and y in the range of the response variable.

(2) The functions $f_{\mathrm{U}}(u)$, $\Gamma(u)$, $V(m(u,\,\mathbf{x}))$, $V'(m(u,\,\mathbf{x}))$ and $g'''(m(u,\,\mathbf{x}))$ are continuous at the point $u = u_0$. Further, assume that $f_{\mathrm{U}}(u_0) > 0$ and $\Gamma(u_0) > 0$.

(3) $K(\cdot)$ has a bounded support.

(4) $a_j''(\cdot)$ is continuous in a neighborhood of u_0 for $j = 1,\,\ldots,\,p$.

(5) $E\left\{|\mathbf{X}|^3 \,|\, U = u\right\}$ is continuous at the point $u = u_0$.

(6) $E(Y^4 \,|\, U = u,\,\mathbf{X} = \mathbf{x})$ is bounded in a neighborhood of $u = u_0$.

Condition (1) guarantees that the local likelihood function (3.1) is concave. It is satisfied by the model (2.1) with a canonical link. Note that Condition (2) implies that $q_1(\cdot,\cdot)$, $q_2(\cdot,\cdot)$, $q_3(\cdot,\cdot)$, $\rho'(\cdot,\cdot)$ and $m'(\cdot,\cdot)$ are continuous.

Proof of Theorem 1:

Recall that $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ maximizes (3.1). Let $\overline{\eta}(u_0,\,u,\,\mathbf{x}) = \overline{\eta}(u_0,\,u,\,x_1,\,\ldots,\,x_p) = \sum_{j=1}^{p}\{a_j(u_0) + a_j'(u_0)(u-u_0)\}\,x_j$, and

$$\boldsymbol{\beta}^* = \gamma_n^{-1}\,\Big(\beta_1 - a_1(u_0),\,\ldots,\,\beta_p - a_p(u_0),\,h\{\beta_{p+1} - a_1'(u_0)\},\,\ldots,\,h\{\beta_{2p} - a_p'(u_0)\}\Big)^T,$$

where $\gamma_n = (n\,h)^{-1/2}$. It can easily be seen that

$$\sum_{j=1}^{p}\{a_j + b_j(U_i - u_0)\}\,X_{ij} = \overline{\eta}(u_0,\,U_i,\,\mathbf{X}_i) + \gamma_n\,\boldsymbol{\beta}^{*T}\,\mathbf{Z}_i,$$

where $\mathbf{Z}_i = \left(\mathbf{X}_i^T,\, (U_i - u_0)/h\, \mathbf{X}_i^T\right)^T$. Then, the local likelihood function $\ell(\boldsymbol{\beta})$ defined in (3.1) becomes

$$\ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left[g^{-1}\left\{\overline{\eta}(u_0,\, U_i,\, \mathbf{X}_i) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right\},\, Y_i\right] K_h(U_i - u_0),$$

which is a function of $\boldsymbol{\beta}^*$, denoted by $\ell(\boldsymbol{\beta}^*)$. Let

$$\widehat{\boldsymbol{\beta}}^* = \gamma_n^{-1}\, \left(\widehat{\beta}_1 - a_1(u_0),\, \ldots,\, \widehat{\beta}_p - a_p(u_0),\, h\left\{\widehat{\beta}_{p+1} - a_1'(u_0)\right\},\, \ldots,\, h\left\{\widehat{\beta}_{2p} - a_p'(u_0)\right\}\right)^T.$$

Then $\widehat{\boldsymbol{\beta}}^*$ maximizes $\ell(\boldsymbol{\beta}^*)$ since $\widehat{\boldsymbol{\beta}}$ maximizes (3.1). Equivalently, $\widehat{\boldsymbol{\beta}}^*$ maximizes the following normalized function

$$\ell_n(\boldsymbol{\beta}^*) = \sum_{i=1}^{n} \left(\ell\left[g^{-1}\left\{\overline{\eta}_i(u_0) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right\},\, Y_i\right] - \ell\left[g^{-1}\left\{\overline{\eta}_i(u_0)\right\},\, Y_i\right]\right) K\left\{(U_i - u_0)/h\right\},$$

where $\overline{\eta}_i(u_0) = \overline{\eta}(u_0,\, U_i,\, \mathbf{X}_i)$.

We remark that Condition (1) implies that $\ell_n(\cdot)$ is concave in $\boldsymbol{\beta}^*$. Using the Taylor expansion of $\ell\left\{g^{-1}(\cdot),\, y\right\}$, we have

$$\begin{aligned}
\ell_n(\boldsymbol{\beta}^*) \;\; = \;\; & \gamma_n \sum_{i=1}^{n} q_1\left\{\overline{\eta}_i(u_0),\, Y_i\right\}\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\, K\left\{(U_i - u_0)/h\right\} \\
& + \frac{\gamma_n^2}{2} \sum_{i=1}^{n} q_2\left\{\overline{\eta}_i(u_0),\, Y_i\right\}\, \left(\boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right)^2\, K\left\{(U_i - u_0)/h\right\} \\
& + \frac{\gamma_n^3}{6} \sum_{i=1}^{n} q_3\left\{\eta_i,\, Y_i\right\}\, \left(\boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right)^3\, K\left\{(U_i - u_0)/h\right\}, \qquad \text{(A.2)}
\end{aligned}$$

where η_i is between $\overline{\eta}_i(u_0)$ and $\overline{\eta}_i(u_0) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i$. Let

$$W_n = \gamma_n \sum_{i=1}^{n} q_1\left\{\overline{\eta}_i(u_0),\, Y_i\right\}\, \mathbf{Z}_i\, K\left\{(U_i - u_0)/h\right\}, \qquad \text{(A.3)}$$

and

$$\Delta_n = \frac{\gamma_n^2}{2} \sum_{i=1}^{n} q_2\left\{\overline{\eta}_i(u_0),\, Y_i\right\}\, \mathbf{Z}_i\, \mathbf{Z}_i^T\, K\left\{(U_i - u_0)/h\right\}.$$

Then, (A.2) becomes

$$\ell_n(\boldsymbol{\beta}^*) = W_n^T\, \boldsymbol{\beta}^* + \frac{1}{2}\, \boldsymbol{\beta}^{*T}\, \Delta_n\, \boldsymbol{\beta}^* + \frac{\gamma_n^3}{6} \sum_{i=1}^{n} q_3\left\{\eta_i,\, Y_i\right\}\, \left(\boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right)^3\, K\left\{(U_i - u_0)/h\right\}. \qquad \text{(A.4)}$$

Note that

$$(\Delta_n)_{ij} = (E\Delta_n)_{ij} + O_p\left[\left\{\mathrm{Var}(\Delta_n)_{ij}\right\}^{1/2}\right].$$

Now the mean in the above expression equals

$$E(\Delta_n) = h^{-1}\, E\left[q_2\left\{\overline{\eta}(u_0,\, U,\, \mathbf{X}),\, m(U,\, \mathbf{X})\right\}\, K\left\{(U - u_0)/h\right\}\, \mathbf{Z}\, \mathbf{Z}^T\right].$$

By a Taylor series expansion of $\eta(u,\, \mathbf{x})$ with respect to u around $|u - u_0| < h$ and the first result in (A.1), we have

$$\eta(u,\, \mathbf{x}) = \overline{\eta}(u_0,\, u,\, \mathbf{x}) + \frac{h^2\,(u-u_0)^2}{2}\,\eta''_u(u_0,\, \mathbf{x}) + o(h^2),$$

where $\eta''_u(u,\, \mathbf{x}) = (\partial^2/\partial u^2)\eta(u,\, \mathbf{x}) = \sum_{j=1}^{p}\, a''_j(u)\, x_j$, which implies that

$$q_1\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\; m(u,\, \mathbf{x})\} = \rho(u,\, \mathbf{x})\,\frac{h^2\,(u-u_0)^2}{2}\,\eta''_u(u_o,\, \mathbf{x}) + o(h^2), \tag{A.5}$$

and

$$q_2\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\; m(u,\, \mathbf{x})\} = -\rho(u,\, \mathbf{x}) + o(1). \tag{A.6}$$

Then, using the second equality of (A.1) and (A.6), we obtain

$$E(\Delta_n) \;\to\; -\,f_{\mathrm{U}}(u_0)\,\begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes \Gamma(u_0) = -\,\Delta, \tag{A.7}$$

where $\Gamma(u_0)$ is given in (3.6) and Δ is defined in (5.2). Similar arguments show that $\mathrm{Var}\{(\Delta_n)_{ij}\} = O\left\{(n\,h)^{-1}\right\}$. Therefore,

$$\Delta_n = -\Delta + o_p(1). \tag{A.8}$$

Since $K(\cdot)$ is bounded, $q_3(\cdot,\, \cdot)$ is linear in Y_1 and $E(|Y_1|\,|\,U_1,\, \mathbf{X}_1) < \infty$, the expected value of the absolute value of the last term in (A.4) is bounded by

$$O\left(n\,\gamma_n^3\,E\Big|q_3(\eta_1,\, Y_1)\,\mathbf{X}_1^3\,K\left\{(U_1-u_0)/h\right\}\Big|\right) = O(\gamma_n)$$

by Condition (5). Therefore, the last term in (A.4) is of order $O_p(\gamma_n)$. This, in conjunction with (A.4), (A.7) and (A.8), implies that

$$\ell_n(\boldsymbol{\beta}^*) = W_n^T\,\boldsymbol{\beta}^* - \frac{1}{2}\,\boldsymbol{\beta}^{*T}\,\Delta\,\boldsymbol{\beta}^* + o_p(1).$$

An application of the quadratic approximation lemma (see, for example, Fan and Gijbel 1996, p.210) leads to

$$\widehat{\boldsymbol{\beta}}^* = \Delta^{-1}\,W_n + o_p(1), \tag{A.9}$$

if W_n is a sequence of stochastically bounded random vectors. The asymptotic normality of $\widehat{\boldsymbol{\beta}}^*$ follows from that of W_n. Hence, it remains to establish the asymptotic normality of W_n.

Note that the random vector W_n is a sum of i.i.d. random vectors. In order to establish its asymptotic normality, it suffices to compute the mean and covariance matrix of W_n and check the Lyapounov condition. To this end, by (A.5), we have

$$\begin{aligned} E(W_n) &= n\,\gamma_n\,E\left[q_1\left\{\overline{\eta}(u_0,\, U,\, \mathbf{X}),\, m(U,\, \mathbf{X})\right\}\,\mathbf{Z}\,K\left\{(U-u_0)/h\right\}\right] \\ &= \frac{h^2\,f_{\mathrm{U}}(u_0)}{2\,\gamma_n}\,\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \Gamma(u_0)\,\mathbf{a}''(u_0)\,\left\{1+o(1)\right\}. \end{aligned} \tag{A.10}$$

Similarly, by (A.10) and the definition of $q_1(\cdot,\, \cdot)$, one has

$$\begin{aligned} \mathrm{Var}(W_n) &= n\,\gamma_n^2\,\mathrm{Var}\left[q_1\left\{\overline{\eta}(u_0,\, U,\, \mathbf{X}),\, Y\right)\right\}\,\mathbf{Z}\,K\left\{(U-u_0)/h\right\}] \\ &= h^{-1}\,E\left[q_1^2\left\{\overline{\eta}(u_0,\, U,\, \mathbf{X}),\, Y\right)\right\}\,\mathbf{Z}\,\mathbf{Z}^T\,K^2\left\{(U-u_0)/h\right\}\right] \\ &= f_{\mathrm{U}}(u_0)\,\begin{pmatrix} \nu_0 & \nu_1 \\ \nu_1 & \nu_2 \end{pmatrix} \otimes \Gamma(u_0)\,\left\{1+o(1)\right\} \\ &= \Lambda + o(1), \end{aligned} \tag{A.11}$$

where Λ is defined in (5.2). We now employ the Cramér-Wold device to derive the asymptotic normality of W_n. For any unit vector $\mathbf{d} \in \Re^{2p}$, if

$$\left\{\mathbf{d}^T \, \mathrm{Var}(W_n)\, \mathbf{d}\right\}^{-1/2} \left\{\mathbf{d}^T \, W_n - \mathbf{d}^T \, E(W_n)\right\} \quad \xrightarrow{\mathcal{D}} \quad N(0,\ 1), \tag{A.12}$$

then

$$\{\mathrm{Var}(W_n)\}^{-1/2} \, (W_n - E(W_n)) \quad \xrightarrow{\mathcal{D}} \quad N(0,\ \mathbf{I}_{2p}). \tag{A.13}$$

Combining (A.9), (A.10), (A.11) and (A.13), we obtain

$$\widehat{\boldsymbol{\beta}}^* - \frac{(n\,h^5)^{1/2}}{2}\, \Delta^{-1}\, f_{\mathrm{U}}(u_0) \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \Gamma(u_0)\, \mathbf{a}''(u_0)\, \{1 + o(1)\} \quad \xrightarrow{\mathcal{D}} \quad N\left(0,\ \Delta^{-1}\, \Lambda\, \Delta^{-1}\right). \tag{A.14}$$

Therefore, the assertion in (5.1) holds true. To prove (A.12), we need only to check Lyapounov's condition for that sequence, which can be easily verified. If $K(\cdot)$ is symmetric, then $\mu_1 = 0$, so that (5.3) holds true. This completes the proof of the theorem. □

Proof of Theorem 2:

For the sake of simplicity, in the process of derivations, some of the notation will be simplified by dropping some of its arguments involved, here and in the sequel. Recall that $\ell(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \ell\left\{g^{-1}\left(\sum_{j=1}^p (a_j + b_j\,(U_i - u_0))\, X_{ij}\right),\ Y_i\right\} K_h(U_i - u_0)$. For any $\widetilde{\boldsymbol{\beta}}$ satisfying $\mathbf{H}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p\left(h^2 + (n\,h)^{-1/2}\right)$, one can easily show that

$$\begin{aligned} \mathbf{H}^{-1}\, \ell''\left(\widetilde{\boldsymbol{\beta}}\right)\, \mathbf{H}^{-1} &= \mathbf{H}^{-1}\, \ell''(\boldsymbol{\beta})\, \mathbf{H}^{-1} + o_p(1) \\ &= \frac{1}{n} \sum_{i=1}^n q_2\left\{\widetilde{\mathbf{Z}}_i^T \boldsymbol{\beta},\ Y_i\right\} \mathbf{H}^{-1}\, \widetilde{\mathbf{Z}}_i\, \widetilde{\mathbf{Z}}_i^T\, \mathbf{H}^{-1}\, K_h(U_i - u_0) + o_p(1), \end{aligned} \tag{A.15}$$

where $\widetilde{\mathbf{Z}}_i = \left(\mathbf{X}_i^T, (U_i - u_0)\mathbf{X}_i^T\right)^T$. By computing the mean and variance of $\mathbf{H}^{-1}\, \ell''(\boldsymbol{\beta})\, \mathbf{H}^{-1}$, we obtain

$$\begin{aligned} & \mathbf{H}^{-1}\, \ell''\left(\widetilde{\boldsymbol{\beta}}\right)\, \mathbf{H}^{-1} \\ = \ & E\left[q_2\left\{\widetilde{\mathbf{Z}}^T \boldsymbol{\beta},\ Y\right\} \begin{pmatrix} 1 & \frac{U-u_0}{h} \\ \frac{U-u_0}{h} & \frac{(U-u_0)^2}{h^2} \end{pmatrix} \otimes \mathbf{X}\,\mathbf{X}^T\, K_h(U - u_0)\right] + o_p(1) \\ = \ & E\left[q_2\left\{\widetilde{\mathbf{Z}}^T \boldsymbol{\beta},\ m(U,\, \mathbf{X})\right\} \begin{pmatrix} 1 & \frac{U-u_0}{h} \\ \frac{U-u_0}{h} & \frac{(U-u_0)^2}{h^2} \end{pmatrix} \otimes \mathbf{X}\,\mathbf{X}^T\, K_h(U - u_0)\right] + o_p(1) \\ = \ & -\Delta + o_p(1), \end{aligned} \tag{A.16}$$

where Δ is defined in (5.2). Recall that $\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} = \widehat{\boldsymbol{\beta}}_0 - \left\{\ell''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \ell'\left(\widehat{\boldsymbol{\beta}}_0\right)$ (see (3.2)). By the Taylor expansion, we have

$$\ell'\left(\widehat{\boldsymbol{\beta}}_0\right) = \ell'(\boldsymbol{\beta}) + \ell''\left(\widetilde{\boldsymbol{\beta}}^*\right)\, \left(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right),$$

where $\widetilde{\boldsymbol{\beta}}^*$ lies between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}_0$ and hence satisfies $\mathbf{H}\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right) = O_p\left(h^2 + (n\,h)^{-1/2}\right)$. Then, some algebraic computations show that

$$\begin{aligned}
\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) &= \mathbf{H}\left(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) - \mathbf{H}\left\{\ell''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \mathbf{H}\,\mathbf{H}^{-1}\,\ell'\left(\widehat{\boldsymbol{\beta}}_0\right) \\
&= \left[\mathbf{I} - \mathbf{H}\left\{\ell''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \mathbf{H}\,\mathbf{H}^{-1}\,\ell''\left(\widetilde{\boldsymbol{\beta}}^*\right)\,\mathbf{H}^{-1}\right]\,\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) \\
&\quad - \mathbf{H}\left\{\ell''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \mathbf{H}\,\mathbf{H}^{-1}\,\ell'(\boldsymbol{\beta}). \qquad \text{(A.17)}
\end{aligned}$$

Therefore, by (A.16) and (A.17), we have

$$\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) = \Delta^{-1}\,\mathbf{H}^{-1}\,\ell'(\boldsymbol{\beta})\,\{1 + o_p(1)\} + o_p\left(h^2 + (n\,h)^{-1/2}\right),$$

which, in conjunction with (A.3), (A.9), (A.13) and (A.14), implies that

$$\sqrt{n\,h}\,\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) = \Delta^{-1}\,W_n + o_p(1) = \widehat{\boldsymbol{\beta}}^* + o_p(1). \qquad \text{(A.18)}$$

Therefore, $\widehat{\boldsymbol{\beta}}_{\mathrm{OS}}$ has the same asymptotic distribution as $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$. □

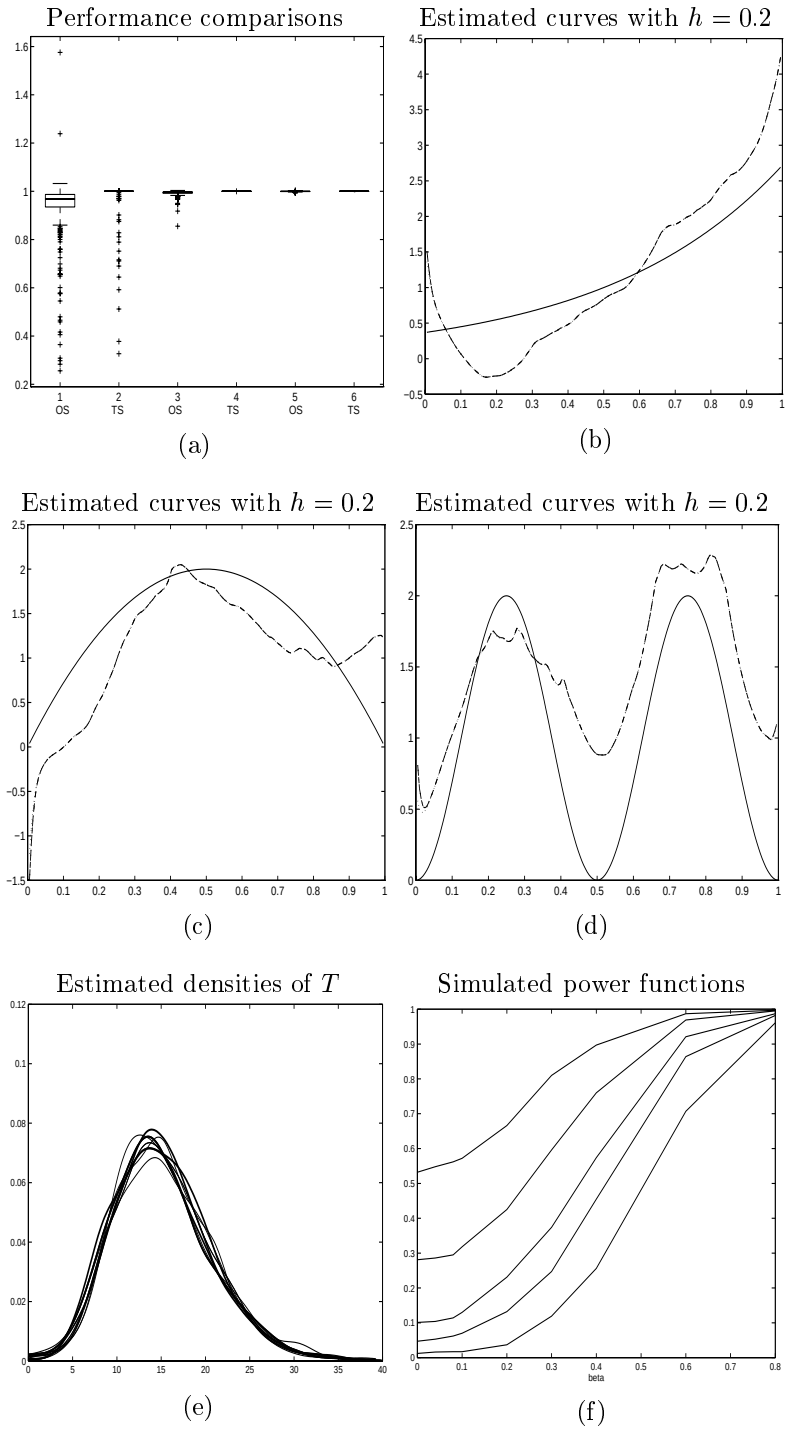


Figure 1: *Simulation results for Example 1 with sample size 400. (a) The boxplots for the ratios of RASE of the one-step and two-step local likelihood approaches to that of the local MLE of* $\mathbf{a}(u)$*, using bandwidths (from left to right)* $h = 0.10$*,* 0.20 *and* 0.40*. (b), (c) and (d) Typical estimates of* $a_0(u)$*,* $a_1(u)$ *and* $a_2(u)$*, respectively, with bandwidth* $h = 0.2$*. Solid curve — true function; dashed curves (from shortest to longest dash) are the one-step, two-step and local MLE, respectively. (e) The estimated densities of* T *for unconditional null distributions (thick solid lines) and for conditional null distributions (thin solid curves). (f) The power functions of the test statistic* T*.*

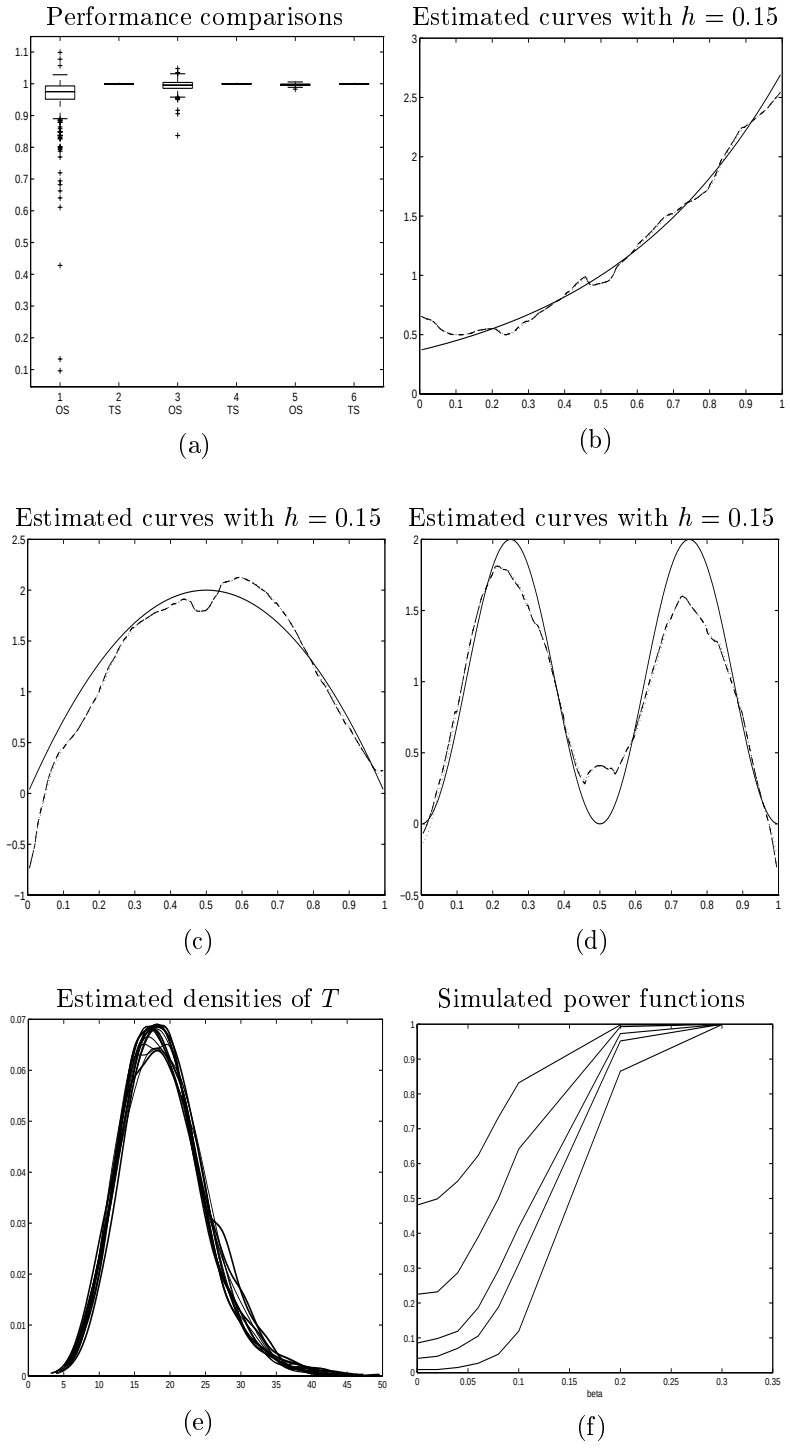
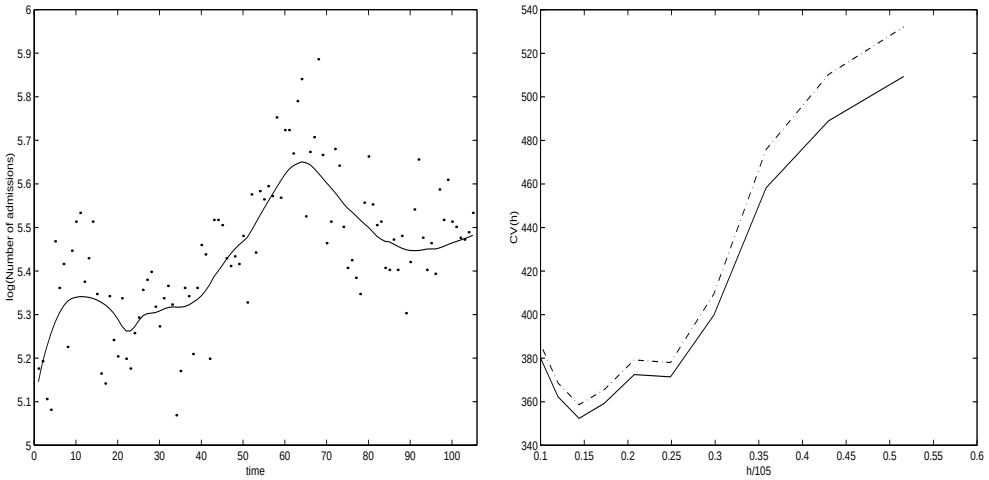


Figure 2: *Simulation results for Example 2 with sample size 200. The caption is similar to Figure 1.*



(a) (b)

Figure 3: *(a) The scatterplot of log transformation of environmental data set studied in §4.3. The curve is the estimate of* $a_1(t) + a_2(t)\,\bar{x}_1 + a_3(t)\,\bar{x}_2 + a_4(t)\,\bar{x}_3$*, where* $\bar{x}_j$ *is the average pollutant level* x_j*. (b) The plot of the cross validation functions* $CV_1(h)$ *(solid line) and* $CV_2(h)$ *(dashdot line) against bandwidth.*

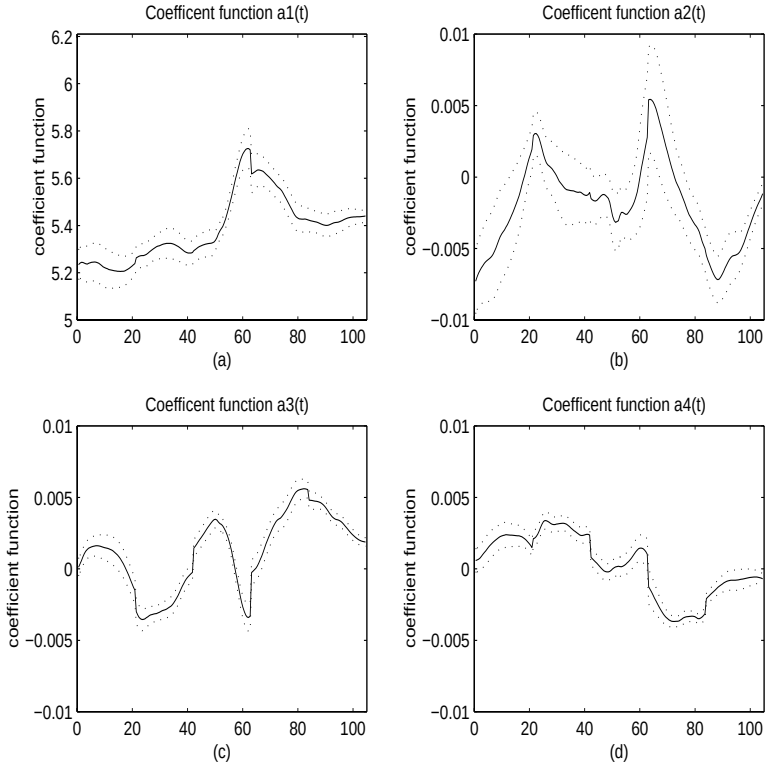
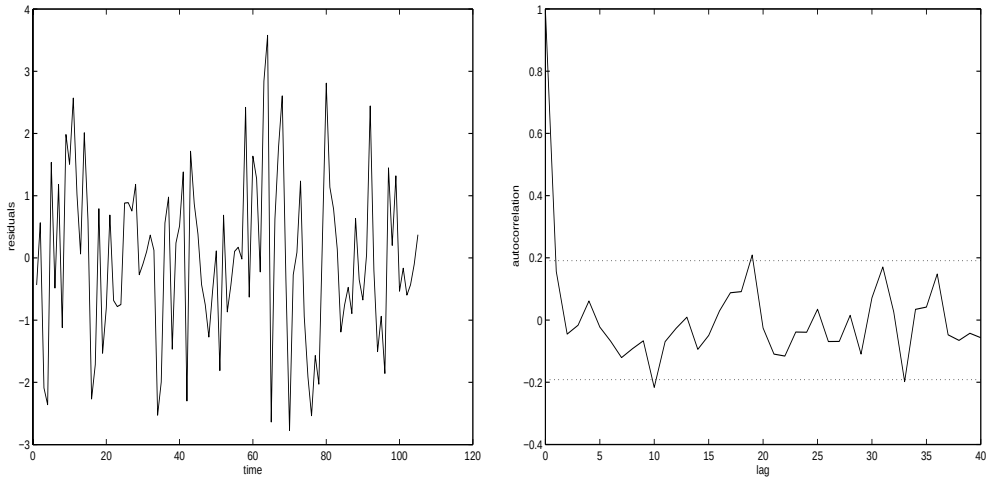


Figure 4: *The estimated coefficient functions via the one-step approach with bandwidth chosen by the CV. The dot curves are the estimated function plus/minus twice estimated standard errors.*



(a) (b)

Figure 5: *(a) The time series plot of Pearson's residuals. (b) The plot of the autocorrelation coefficients versus time lag. The two dot lines are* $\pm 1.96/\sqrt{n}$*, where* n *is the sample size.*

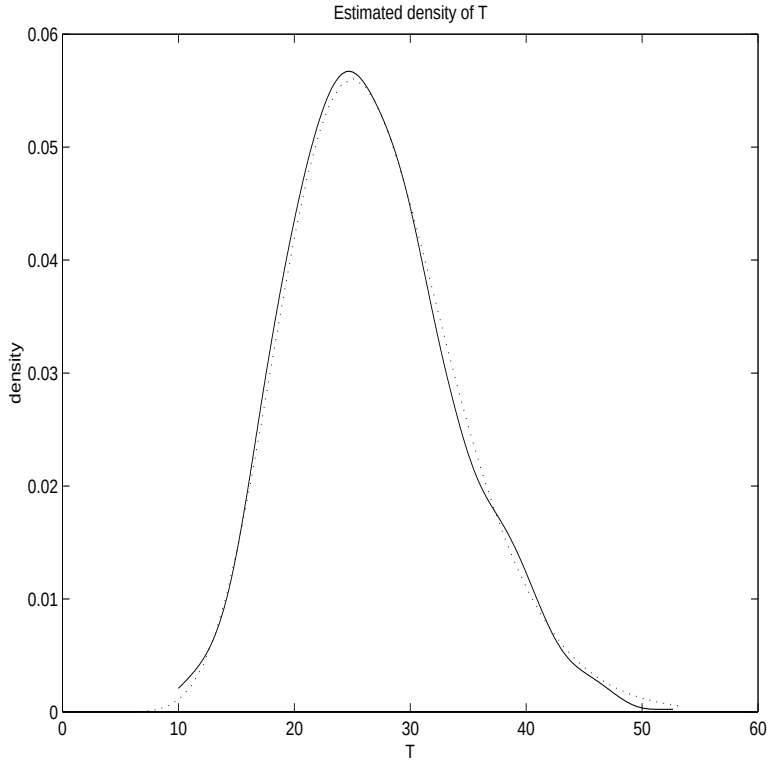


Figure 6: *The estimated density of* T *by Monte Carlo simulation. The solid curve is the estimated density, and the dot curve stands for the density of chi-squared distribution with degrees of freedom 27.*

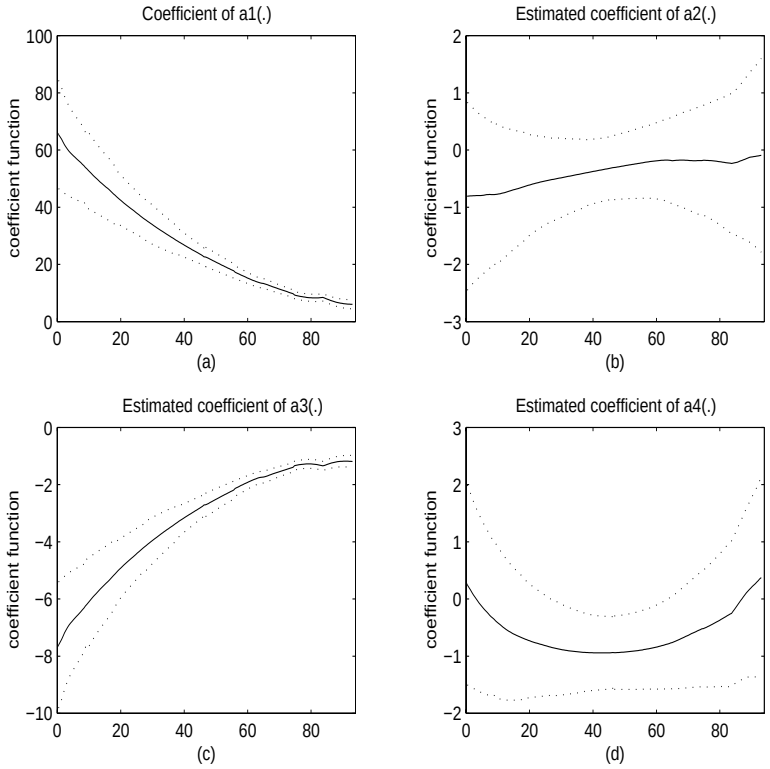


Figure 7: *The estimated coefficient functions (the solid curves) via one-step approach with bandwidth chosen by the CV. The dot curves are the estimated functions plus/minus twice estimated standard errors.*