Efficient Estimation and Inferences for Varying-Coefficient Models [*]

Zongwu Cai   Jianqing Fan   and   Runze Li

Abstract

This paper deals with statistical inferences based on the 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 computational 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.

[*] Zongwu Cai is Assistant Professor, Department of Mathematics, University of North Carolina, Charlotte, NC 28223. E-mail: zcai@uncc.edu. Jianqing Fan is Professor, Department of Statistics, University of California, Los Angeles, CA 90095. E-mail: jfan@math.ucla.edu. Runze Li is Graduate Student, Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. E-mail: lirz@email.unc.edu. Fan's research was partially supported by NSF grant DMS-9803200. We would like to thank the Editor, the Associate Editor and two referees for their constructive and detailed suggestions that led to improving significantly the presentation of the paper.

1 Introduction

Generalized linear models 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. In recent years, various attempts have been made to relax these model assumptions and hence widen their applicability, since a wrong model on the regression function can lead to excessive modeling biases and erroneous conclusions. Of importance is the varying-coefficient models, proposed by Hastie and Tibshirani (1993), which widen the scope of applications by allowing regression coefficients to depend on certain covariates.

A varying-coefficient model has the form

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

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 $\mathbf{X} = \mathbf{x}$ with $\mathbf{X} = (X_1, \ldots, X_p)^T$. Clearly, model (1.1) includes both the parametric generalized linear model (McCullagh and Nelder 1989) and the generalized partially linear model (Chen 1988; Speckman 1988; Green and Silverman 1994; Carroll, Fan, Gijbels and Wand 1997).

A motivation of this study comes from an analysis of environmental data, consisting of weekly measurements of pollutants and other environmental factors, collected in Hong Kong from January 1, 1994 to December 31, 1995 (Courtesy of Professor T. S. Lau). Of interest is to examine the association between the levels of pollutants and the total number of weekly hospital admissions for circulatory and respiratory problems. It is natural to allow the association to change over time (see Figure 3(a) below). Such a problem can be tackled by using model (1.1) as follows. 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 is modeled as a Poisson distribution with the mean function given by (1.1). 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. Detailed analyses of these two data sets will be reported in §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. Recently, some new developement has been made to the model (1.1). Kauermann and Tutz (1999) proposed a graphical procedure to diagnose the discrepancy between the parametric model and the smoothing alternative by using the local likelihood smoothing. Furthermore, a two-step estimation procedure was proposed by Fan and Zhang (2000) to deal with the situations where coefficient functions admit different degrees of smoothness. An advantage of model (1.1) is that by allowing the coefficients $\{a_j(\cdot)\}$ to depend on $\mathbf{U}$, the modeling bias can be reduced significantly and the "curse of dimensionality" is avoided.

Varying-coefficient models are a simple and useful extension of classical generalized linear models. This extension admits simple interpretation. 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 (2000) 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 the functional-coefficient autoregressive models. Cai, Fan and Yao (1998) gave an extensive study on the advantages of the varying-coefficient model over the parametric model based on the predictive utility. For applications in finance and econometrics, we refer to the unpublished papers by Hong and Lee (1999) and Lee and Ullah (1999).

Estimation of the coefficient functions in (1.1) is obtained by using local smoothing techniques. By localizing data around $\mathbf{u}$, model (1.1) is approximately a generalized linear model. One can find its local maximum likelihood estimate (MLE) using an iterative algorithm. Note that the local MLE for the varying-coefficient model is indeed solving the local likelihood equations. Thus, our local likelihood method can be regarded as a special case of the general local estimation equation method proposed by Carroll, Ruppert and Welsh (1998). Hence, the bandwidth involved can be selected by the empirical bias method proposed in that paper. In order to obtain the estimated coefficient functions, we need to solve hundreds of local maximum likelihood problems. The computation can be expensive, depending on the convergence criterion. Computational burden becomes even more severe when a cross-validation method is used to select a smoothing parameter. To reduce computational costs, we propose a one-step local MLE. The idea is not novel since it was first used by Bickel (1975) in the parametric setting, but implementations and insights are. We will show that computational costs can be reduced significantly 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 the varying-coefficient models are the standard errors of the estimated coefficient functions. Consistent estimates are derived. Our simulation studies show that the estimated standard errors are very accurate for most applications. Another important issue arises regarding whether some of the coefficient functions in model (1.1) are actually varying, 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 performs well.

One of our goals is to estimate efficiently the coefficient functions $\{a_j(\cdot)\}$ in model (1.1) by using a nonparametric method. Our methods are directly applicable to situations in which one can not specify fully the conditional log-likelihood function $\ell(v, y)$, but can model the relationship between the mean and variance by $\mathrm{Var}(Y \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x}) = \sigma^2 V\{m(\mathbf{u}, \mathbf{x})\}$ for a known variance function $V(\cdot)$ and unknown σ. In this case, one needs only to replace the log-likelihood function $\ell(v, y)$ by the quasi-likelihood function $Q(\cdot, \cdot)$, defined by $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$. It is assumed throughout this paper that the conditional log-likelihood function $\ell(v, y)$ is known and linear in y for fixed v. This assumption is satisfied for the canonical exponential family, which is the focus of this paper.

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

2 Modeling Procedures

For simplicity, we consider only the case that $\mathbf{u}$ is one-dimensional. Extension to multivariate $\mathbf{u}$ involves no fundamentally new ideas. However, implementations with $\mathbf{u}$ having more than two dimensions may have some difficulties due to the "curse of dimensionality."

2.1 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), design adaptation (Fan 1993) and good boundary behavior (Ruppert and Wand 1994; Fan and Gijbels 1996). Suppose that $a_j(\cdot)$ has a continuous second derivative. For each given point u_0, we approximate $a_j(u)$ locally by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$ for u in a neighborhood of u_0. Based on a random sample $\{(U_i,\, \mathbf{X}_i,\, Y_i)\}_{i=1}^{n}$, we use the following local likelihood method to estimate the coefficient functions

$$\ell_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left[g^{-1}\left\{\sum_{j=1}^{p}(a_j + b_j(U_i - u_0))X_{ij}\right\}, Y_i\right] K_h(U_i - u_0), \tag{2.1}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, $h = h_n > 0$ is a bandwidth, $\mathbf{a} = (a_1,\, \ldots,\, a_p)^T$ and $\mathbf{b} = (b_1,\, \ldots,\, b_p)^T$. Note that a_j and b_j are dependent on u_0, and so is $\ell_n(\cdot,\, \cdot)$. Maximizing the local likelihood function $\ell_n(\mathbf{a},\, \mathbf{b})$ 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 denote $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1,\, \ldots,\, a_p,\, b_1,\, \ldots,\, b_p)^T$ and write the local likelihood function (2.1) as $\ell_n(\boldsymbol{\beta})$. Likewise, the local MLE is denoted by $\widehat{\boldsymbol{\beta}}_{\text{MLE}} = \widehat{\boldsymbol{\beta}}_{\text{MLE}}(u_0)$.

2.2 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 the estimated functions $\{\widehat{a}_j(\cdot)\}$, one needs to maximize the local likelihood (2.1) for usually hundreds of distinct values of u_0, with each maximization requiring an iterative algorithm. Moreover, the computational expense further increases with the number of covariates p. To ameliorate this expense, we propose to replace the iterative local MLE by the one-step Newton-Raphson estimator, which has been frequently used in parametric models (Bickel 1975; Lehmann 1983). We prove Theorem 2 below that the one-step local MLE does not lose any statistical efficiency provided that the initial estimator is good enough.

Let $\ell_n'(\boldsymbol{\beta})$ and $\ell_n''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local log-likelihood $\ell_n(\boldsymbol{\beta})$. Given an initial estimator $\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}}_0(u_0) = \left(\widehat{\mathbf{a}}(u_0)^T,\, \widehat{\mathbf{b}}(u_0)^T\right)^T$, one-step of the Newton-Raphson algorithm produces the updated estimator,

$$\widehat{\boldsymbol{\beta}}_{\text{OS}} = \widehat{\boldsymbol{\beta}}_0 - \left\{\ell_n''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \ell_n'\left(\widehat{\boldsymbol{\beta}}_0\right), \tag{2.2}$$

thus featuring the computational expediency of least-squares local polynomial fitting. In univariate generalized linear models, Fan and Chen (1999) carefully studied properties of the local one-step estimator. In that setting, the least-squares estimate serves a natural candidate as an initial estimator, however, in the multivariate setting, it is not clear how an initial estimator can be constructed.

Note that $\ell_n''(\widehat{\boldsymbol\beta}_0)$ can be nearly singular for certain u_0, due to possible data sparsity in certain local regions, or when the bandwidth is too small. Seifert and Gasser (1996) and Fan and Chen (1999) explored the use of the ridge regression as an approach to handling such problems in the univariate setting. We extend their ideas in §3.

2.3 Sampling properties

We now 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.

Define $\mu_k = \int u^k\, K(u)\, d\, u$ and $\nu_k = \int u^k\, K^2(u)\, du$. Let $\mathbf{H} = {\rm diag}(1,\, h) \otimes \mathbf{I}_p$ with $\otimes$ denoting the Kronecker product. Let $f_{\rm U}(\cdot)$ denote the marginal density of U,

$$\Gamma(u) = E\left\{\rho(U,\, \mathbf{X})\, \mathbf{X}\, \mathbf{X}^T \,|\, U = u\right\}, \tag{2.3}$$

and

$$\rho(u,\, \mathbf{x}) = [g_1\{m(u,\, \mathbf{x})\}]^2\ {\rm Var}\{Y|U = u,\, \mathbf{X} = \mathbf{x}\} \tag{2.4}$$

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 exponential family with the canonical link function. The asymptotic properties of $\widehat{\boldsymbol\beta}_{\rm MLE}$ and $\widehat{\boldsymbol\beta}_{\rm OS}$ are described in the following theorems, with conditions and proofs discussed 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

$$\begin{aligned}&\sqrt{n\, h}\left[\mathbf{H}\left\{\widehat{\boldsymbol\beta}_{\rm MLE}(u_0) - \boldsymbol\beta(u_0)\right\} - \frac{h^2}{2(\mu_2 - \mu_1^2)}\begin{pmatrix}(\mu_2^2 - \mu_1\, \mu_3)\, \mathbf{a}''(u_0)\\ (\mu_3 - \mu_1\, \mu_2)\, \mathbf{a}''(u_0)\end{pmatrix} + o_p(h^2)\right]\\ &\xrightarrow{\mathcal{D}}\quad N\left(0,\ \Delta^{-1}\, \Lambda\, \Delta^{-1}\right)\end{aligned} \tag{2.5}$$

with $\Gamma(u_0)$ given by (2.3),

$$\Delta = f_{\rm U}(u_0)\begin{pmatrix}1 & \mu_1\\ \mu_1 & \mu_2\end{pmatrix} \otimes \Gamma(u_0) \quad \text{and} \quad \Lambda = f_{\rm U}(u_0)\begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix} \otimes \Gamma(u_0). \tag{2.6}$$

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

$$\sqrt{n\,h}\,\left[\widehat{\mathbf{a}}_{\mathrm{MLE}}(u_0) - \mathbf{a}(u_0) - \frac{h^2\,\mu_2}{2}\,\mathbf{a}''(u_0) + o_p(h^2)\right] \quad \xrightarrow{\mathcal{D}} \quad N\,(0,\ \Sigma(u_0))\,, \tag{2.7}$$

where

$$\Sigma(u_0) = \nu_0\,\Gamma^{-1}(u_0)/f_{\mathrm{U}}(u_0). \tag{2.8}$$

Note that the bias and variance expressions in Theorem 1 can be deduced from the general theorem from Carroll, Rupport and Welsh (1998). However, the main difference here is that we establish the results in terms of asymptotic normality, while they established them for the general case using conditional expections.

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

$$\mathbf{H}\,\left(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) = O_p\left\{h^2 + (n\,h)^{-1/2}\right\}. \tag{2.9}$$

As a consequence, the fully iterative MLE and the one-step estimate share the same asymptotic properties provided that (2.9) is fulfilled, which provide the theoretical basis for the use of the one-step approach in practice. The asymptotic mean squared error (MSE) of the two estimators $\widehat{a}_{j,\mathrm{MLE}}(u_0)$ and $\widehat{a}_{j,\mathrm{OS}}(u_0)$ is

$$\mathrm{MSE} = \frac{h^4}{4}\,\mu_2^2\,\left\{a_j''(u_0)\right\}^2 + \frac{\nu_0\,\sigma_{jj}^2(u_0)}{n\,h\,f_U(u_0)},$$

when $K(\cdot)$ is symmetric, where $\sigma_{jj}^2(u_0)$ is the j-th diagonal element of $\Gamma^{-1}(u_0)$. Then, the MSE is of order $n^{-4/5}$ if the optimal bandwidth $[\nu_0\,\sigma_{jj}^2(u_0)/\{\mu_2^2\,(a_j''(u_o))^2\,f_U(u_0)\}]^{1/5}\,n^{-1/5}$ is used.

2.4 Standard errors

Since the local likelihood (2.1) is a weighted likelihood function of a parametric generalized linear model, the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ can be estimated from conventional techniques. Let $q_j(s,\,y) = (\partial^j/\partial s^j)\,\ell\,\{g^{-1}(s),\,y\}$ and

$$\widehat{\Gamma}(u_0) = -\,\frac{1}{n}\sum_{i=1}^{n} q_2\left[\sum_{j=1}^{p}\{\widehat{a}_j(u_0)\,X_{ij} + \widehat{b}_j(u_0)(U_i - u_0)\},\ Y_i\right]\,K_h(U_i - u_0)\begin{pmatrix}\mathbf{X}_i\\ \mathbf{X}_i\,(U_i - u_0)/h\end{pmatrix}^{\otimes 2}, \tag{2.10}$$

where $A^{\otimes 2}$ denotes $A\,A^T$ for a matrix or vector A. Then, the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ can be estimated as

$$\widehat{\Sigma}^*(u_0) = \widehat{\Gamma}(u_0)^{-1}\,\widehat{\Lambda}(u_0)\,\widehat{\Gamma}(u_0)^{-1}, \tag{2.11}$$

where

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

In the implementation in §3, a ridge regression technique is employed and hence the matrix $\widehat{\Gamma}(u_0)$ in (2.11) is slightly modified to reflect this change.

The explicit formula for the asymptotic covariance matrix in (2.8) provides an alternative estimate of the asymptotic covariance matrix of $\mathbf{a}(u_0)$ (not full vector $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$) $\Sigma(u_0)$. Therefore, a direct estimate of $\Sigma(u_0)$ is $\widetilde{\Sigma}(u_0) = \nu_0\,\widehat{\Gamma}_S(u_0)^{-1}$, where $\widehat{\Gamma}_S(u_0)$ is the $p\times p$ upper corner submatrix of $\widehat{\Gamma}(u_0)$ given by (2.10).

2.5 Hypothesis testing

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

$$H_0 : a_1(u) \equiv a_1,\ \cdots,\ a_p(u) \equiv a_p, \tag{2.12}$$

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

$$T = 2\{\ell(H_1) - \ell(H_0)\}, \tag{2.13}$$

where $\ell(H_0)$ and $\ell(H_1)$ are respectively the log-likelihood functions computed under the null and alternative hypotheses.

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 $a_1, \cdots, a_p$. For the rigorous justification, we refer to the paper by Fan, Zhang and Zhang (1999) who considered sieve likelihood ratio tests in a general setting and demonstrated that the Wilks' type of phenomenon holds for a large variety of nonparametric problems. This in turn suggests that we can use the following *conditional bootstrap* to construct the null distribution of T. Let $\{\widehat{a}_j\}$ be the MLE under the null hypothesis. Given the covariates $(U_i, \mathbf{X}_i)$, generate a bootstrap sample Y_i^* from the given distribution of Y with the estimated linear predictor $\widehat{\eta}(U_i, \mathbf{X}_i) = \sum_{j=1}^{p}\widehat{a}_j X_{ij}$ and compute the test statistic T^* in (2.13). 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 $\{a_j\}$ (Fan, Zhang, and Zhang, 1999).

Note that the above conditional bootstrap method applies readily to the Poisson and Bernoulli distributions, since in these cases the distribution of Y does not involve 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 out above, this approximation is valid under both H_0 and H_1 as the null distribution does not asymptotically depend on the values of $\{a_j\}$. 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. This is again due to the Wilks type of results demonstrated by Fan, Zhang and Zhang (1999).

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.

2.6 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 (2.2) 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 (2.2) 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 (2.2) to obtain $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{61})$. Regarding $\widehat{\boldsymbol{\beta}}_{\text{OS}}(u_{61})$ as an initial value, we use (2.2) 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 (2.2) 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 §3 for details.

The theoretical basis for the above "one-step" and the "two-step" procedures is as follows. When the grid points are sufficiently 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.

Based on the above theoretical considerations, we suggest a very simple rule of thumb for choosing the number of grid points: $n_{\text{grid}} = \max\{200, \text{IQR}^2/h^2\}$, where IQR is the interquantile range of $U_1, \cdots, U_n$. In such a way, approximation errors between estimates at two consecutive grid points are of order $O(h^2)$, satisfying the critical condition (2.9).

3 Simulations and Applications

In this section, we first discuss how to implement the one-step procedure for the Bernoulli and 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 $\widehat{\mathbf{a}}(\cdot)$ is assessed via the square-Root of Average Square Errors (RASE)

$$\text{RASE}^2 = n_{\text{grid}}^{-1} \sum_{j=1}^{p} \sum_{k=1}^{n_{\text{grid}}} \{\widehat{a}_j(u_k) - a_j(u_k)\}^2, \tag{3.1}$$

where $\{u_k,\, k = 1, \ldots, n_{grid}\}$ are the grid points at which the functions $\{a_j(\cdot)\}$ are estimated.

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 to represent widely varying degrees of smoothness. Over this 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 $n_{\text{grid}} = 200$ are used.

3.1 Logistic Regression

For a Bernoulli distribution, the one-step estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\text{OS}} = \widehat{\boldsymbol{\beta}}_0 + \begin{pmatrix} \mathbf{H}_{n,0}, & \mathbf{H}_{n,1} \\ \mathbf{H}_{n,1}, & \mathbf{H}_{n,2} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{v}_{n,0} \\ \mathbf{v}_{n,1} \end{pmatrix}, \tag{3.2}$$

where $\mathbf{H}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)\widehat{p}_{i0}(1-\widehat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i\mathbf{X}_i^T$, $j = 0,\, 1,\, 2$, $\widehat{p}_{i0}$ satisfies $\operatorname{logit}(\widehat{p}_{i0}) = \sum_{j=1}^p \left\{\widehat{a}_{j,0} + \widehat{b}_{j,0}(U_i - u_0)\right\} X_{ij}$, and $\mathbf{v}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)(Y_i - \widehat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i$, $j = 0,\, 1$. The two-step estimator $\widehat{\boldsymbol{\beta}}_{\text{TS}}$ is obtained by iterating the equation (3.2) twice and the local MLE is obtained by iterating equation (3.2) until convergence.

In practice, the matrix in (3.2) can be singular or nearly singular when the local data are sparse. To attenuate this difficulty, one may follow the idea of ridge regression (Seifert and Gasser 1996; 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

$$E\left(X_k^2 \,|\, U = u_0\right)\, \widehat{p}_0(1-\widehat{p}_0)\, h^{j-1} \int u^j K(u)\, du\, N \qquad \text{with} \qquad \widehat{p}_0 = \frac{\exp(\widehat{\mathbf{a}}_0^T \overline{\mathbf{X}})}{1+\exp(\widehat{\mathbf{a}}_0^T \overline{\mathbf{X}})}, \tag{3.3}$$

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 becoming nearly singular when N is small. However, it affects and indeed ameliorates the finite-sample properties of estimators for small-sample sizes.

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

$$a_0(u) = \exp(2u-1), \quad a_1(u) = 8u(1-u), \quad \text{and} \quad a_2(u) = 2\,\sin^2(2\pi u). \tag{3.4}$$

Figure 1(a) depicts the marginal distributions for the ratios of the overall RASE defined in (3.1), using three bandwidths $h = 0.1$, 0.2 and 0.4. 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 the estimate of the coefficient functions from a typical sample. The typical sample is selected in such a way that its RASE-value is the median in the 400 RASE-values. Table 1 summarizes the simulation results with μ and σ denoting the mean and standard deviation 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.

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_*
	0.10	2.2278	2.0874	1.8537	0.9759	0.8656	2.1244	1.5315	0.8274
400	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
	0.075	1.2451	0.6639	1.1644	0.3767	0.8342	1.2256	0.5301	0.9656
800	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

We now test the accuracy of our standard error formula (2.11). 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 (2.11). 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 Carlo 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.

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)

Next, we conduct a simulation study to see whether the asymptotic null distribution of the test statistic T defined in (2.13) depends on the values of $\{a_j\}$ under H_0 (see (2.12)) and the limiting conditional null distributions are dependent on the covariate values. 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 very close, which suggest that the asymptotic null distribution is not very sensitive to 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 are very close to the true null distribution. This demonstrates empirically that our bootstrap method gives a reasonably good approximation to the true null distribution even when the data were generated from an alternative model (3.4).

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

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

$$H_0 : a_j(u) = \theta_j, \quad j = 0, 1, 2, \quad \text{versus} \quad H_1 : a_j(u) \neq \theta_j, \text{ for at least one } j.$$

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

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

where $\{a_j^0(u)\}$ are given in (3.4) 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 significance 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 five significance 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 §2.5 works well.

3.2 Poisson regression

For a Poisson model with the canonical link, by straightforward calculation, the one-step estimator is given similarly to (3.2) but now $\mathbf{H}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)\widehat{\lambda}_{i0}(U_i - u_0)^j \mathbf{X}_i \mathbf{X}_i^T$, $j = 0,\, 1,\, 2$, $\widehat{\lambda}_{i0} = \exp\left[\sum_{j=1}^p \{\widehat{a}_{j0} + \widehat{b}_{j0}(U_i - u_0)\} X_{ij}\right]$, and $\mathbf{v}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)(Y_i - \widehat{\lambda}_{i0})(U_i - u_0)^j \mathbf{X}_i$, $j = 0,\, 1$. Using the same arguments as in the previous section, the ridge parameters

$$r_{j,k} = \left(\frac{1}{n}\sum_{i=1}^n X_{ik}^2\right) \widehat{\lambda}_0\, h^{j-1} \int u^j K(u)\, du \qquad \text{with} \qquad \widehat{\lambda}_0 = \exp\left(\widehat{\mathbf{a}}_0^T \overline{\mathbf{X}}\right) \tag{3.5}$$

are employed to alleviate the possible singularity of matrix $\mathbf{H}_{n,j}$ ($j = 0$ and 2) in (3.2).

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 §3.3.

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

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

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 (2.11) in Table 5. Clearly, our estimated standard errors are very close to the true ones.

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).

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

The same conclusions as those in Example 1 can be drawn for the Poisson regression model. In particular, the test has the correct levels of significance. See the power functions in Figure 2(e) at $\beta = 0$.

3.3 Real-data examples

Example 3. We illustrate in this example 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 circulatory and respiratory 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 $\mu g/m^3$). Since the admissions "events" occur at certain points in time, it is reasonable to model the number of admissions as a Poisson process and use the Poisson regression model with the mean $\lambda(t, \mathbf{x})$ given by

$$\log\{\lambda(t, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2 + a_3(t)\, x_3 + a_4(t)\, x_4. \tag{3.6}$$

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

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

For each j, the j-th group of data is deleted and model (3.6) is fitted for the remaining data. Then the deviance (McCullagh and Nelder 1989, p.34) or the sum of squares of Pearson's residuals is computed. This leads to two cross-validation criteria

$$CV_1(h) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} 2 \left[y_i \log\{y_i / \widehat{y}_{-d_j}(U_i, \mathbf{X}_i)\} - \{y_i - \widehat{y}_{-d_j}(U_i, \mathbf{X}_i)\} \right],$$

and

$$CV_2(h) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left\{ \frac{y_i - \widehat{y}_{-d_j}(U_i, \mathbf{X}_i)}{\sqrt{\widehat{y}_{-d_j}(U_i, \mathbf{X}_i)}} \right\}^2,$$

where $\widehat{y}_{-d_j}(U_i, \mathbf{X}_i)$ is a fitted value with the data in d_j deleted. In the implementation, we choose $Q = 20$. Figure 3(b) depicts the cross-validation functions $CV_1(h)$ and $CV_2(h)$ which give the optimal bandwidth $h = 0.1440 \times 105$. To see how sensitive the above partition is to the CV curves, the data set is randomly partitioned into 20 groups, and then cross-validation scores are computed based on the same procedure described above. The results are depicted in Figure 3(c), which, in conjunction with Figure 3(b), shows that the cross-validation functions is not very sensitive to the partition. The estimated coefficient functions based on the one-step procedure are summarized in Figure 4 since the results based on both the one-step and the fully iterative methods are very close. 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 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). Thus, Figure 5(a) together with Figure 5(b) lead to the conclusion that there is no evidence that the data are serially correlated.

We now apply the procedure proposed in §2.5 to testing whether the coefficients are actually 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 (2.13) is $T = 389.41$. Based on 1000 bootstrap replications, the sample mean and sample variance of T^* are 26.64 and 48.40, respectively. The distribution of T is approximated by a χ^2 distribution with degrees of freedom 27 (see Figure 6). The p-value is close to zero, which strongly rejects the null hypothesis. Therefore, it suggests that the varying-coefficient model gives a much better fit than the parametric model.

Now we use our testing approach proposed in §2.5 to check whether there is any covariate that can be deleted from the model. We start with X_4 since the parametric Poisson model concludes that the dust level (X_4) is not statistically significant. To examine if the variable X_4 is significant in the varying-coefficient model, we apply the idea in §2.5 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 can be dropped from the varying-coefficient model. After deleting the variable dust level (X_4), we apply the same procedure as above to test whether X_3 is statistically significant in the 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 the significant level 0.05. By the same token, the variable sulfur dioxide (X_2) is significant too.

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(burn area+1) and binary variable X_4 =*Oxygen* (0 if oxygen supply is normal, 1 otherwise) are considered. We are interested in studying how burn areas and the other variables affect survival probabilities for victims at different age groups. This naturally leads to the following varying-coefficient model

$$\text{logit}\{p(x_1, x_2, x_3, x_4)\} = a_1(x_1) + a_2(x_1)\, x_2 + a_3(x_1)\, x_3 + a_4(x_1)\, x_4. \tag{3.7}$$

Figure 7 presents the estimated coefficients for model (3.7) via the one-step approach with bandwidth $h = 65.7882$, selected by a cross-validation method.

A natural question arises whether the coefficients in (3.7) are actually varying. To see this, we consider the parametric logistic regression model

$$\text{logit}\{p(x_1,\, x_2,\, x_3,\, x_4)\} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \beta_4\, x_4 \tag{3.8}$$

as the null model. As a result, the MLE of $(\beta_0, \cdots, \beta_4)$ in model (3.8) and its standard deviation 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 §2.5 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.

To examine whether there is any gender gap for different age groups or if the variable X_4 affects the survival probabilities for different age of burn victims, we consider testing hypothesis H_0 : both $a_2(\cdot)$ and $a_4(\cdot)$ are constant under model (3.7). 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

(3.7). 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 suggests that gender and oxygen do not play a significant role in determining the survival probability of a victim.

Appendix: Proofs

Before we present the proofs of the theorems, we first impose some regularity conditions. To this end, let us recall that $q_j(s,\, y) = (\partial^j / \partial s^j)\, \ell\, \{g^{-1}(s),\, y\}$. Note that $q_k(s,\, y)$ is linear in y for fixed s such that

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

where $\rho(u,\, \mathbf{x})$ is defined in (2.4). Note that we use the same notation as in §2.

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 (2.1) is concave. It is satisfied for the canonical exponential family 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 (2.1). Let $\overline{\eta}(u_0,\, u,\, \mathbf{x}) = \sum_{j=1}^{p}\{a_j(u_0) + a_j'(u_0)(u - u_0)\}\, x_j$, and

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

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 $\ell_n(\boldsymbol{\beta})$ defined in (2.1) becomes

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

which is a function of $\boldsymbol{\beta}^*$, denoted by $\ell_n(\boldsymbol{\beta}^*)$. 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 $\ell_n(\boldsymbol{\beta}^*)$ since $\widehat{\boldsymbol{\beta}}$ maximizes (2.1). 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

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

where

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

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

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

$$q_1\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\; m(u,\, \mathbf{x})\} = \rho(u,\, \mathbf{x})\, \frac{h^2\,(u-u_0)^2}{2}\, \eta_u''(u_0,\, \mathbf{x}) + o(h^2), \tag{A.4}$$

and

$$q_2\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\; m(u,\, \mathbf{x})\} = -\rho(u,\, \mathbf{x}) + o(1). \tag{A.5}$$

Then, using the second equality of (A.1) and (A.5), we obtain

$$E(\Delta_n) \;\rightarrow\; -\, f_{\mathrm{U}}(u_0) \begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes \Gamma(u_0) = -\,\Delta, \tag{A.6}$$

where $\Gamma(u_0)$ is given in (2.3) and Δ is defined in (2.6). Similar arguments show that $\mathrm{Var}\{(\Delta_n)_{ij}\} = O\left\{(n\,h)^{-1}\right\}$. Therefore,

$$\Delta_n = -\Delta + o_p(1). \tag{A.7}$$

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.2) is bounded by

$$O\left(n\,\gamma_n^3\, E\Big|q_3(\eta_1,\, Y_1)\, \mathbf{X}_1^3\, K\left\{(U_1-u_0)/h\right\}\Big|\right) = O(\gamma_n) \tag{A.8}$$

by Condition (5). Therefore, the last term in (A.2) is of order $O_p(\gamma_n)$. This, in conjunction with (A.2), (A.6) and (A.7), 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.4), 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

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

where Λ is defined in (2.6). By the Cramér-Wold device, in order to derive the asymptotic normality of W_n, it suffices to show that for any unit vector $\mathbf{d} \in \Re^{2p}$,

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

This, conjunction with (A.9), (A.10), and (A.11), implies that

$$\widehat{\boldsymbol{\beta}}^* - \frac{(n\, h^5)^{1/2}}{2}\, \Delta^{-1}\, f_{\mathrm{U}}(u_0)\begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \otimes \Gamma(u_0)\, \mathbf{a}''(u_0)\, \{1+o(1)\} \quad \xrightarrow{\mathcal{D}} \quad N\left(0,\, \Delta^{-1}\, \Lambda\, \Delta^{-1}\right). \tag{A.13}$$

Therefore, the assertion in (2.5) holds true. To prove (A.12), we need only to check Lyapounov's condition for that sequence. To do so, let $\xi_i = q_1\, \{\overline{\eta}_i(u_0),\, Y_i\}\, \mathbf{d}^T\, \mathbf{Z}_i\, K\, \{(U_i - u_0)/h\}$. Then, $\mathbf{d}^T\, W_n = \gamma_n \sum_{i=1}^n \xi_i$. It suffices to show that $n\, \gamma_n^3\, E|\xi_1|^3 \to 0$ as $n \to \infty$. Similar to (A.8), one can show that $n\, \gamma_n^3\, E|\xi_1|^3 = O(\gamma) \to 0$. If $K(\cdot)$ is symmetric, then $\mu_1 = 0$, so that (2.7) holds true. This completes the proof of the theorem. □

Proof of Theorem 2: Recall that

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \ell\left\{g^{-1}\left(\sum_{j=1}^p (a_j + b_j\, (U_i - u_0))\, X_{ij}\right),\, Y_i\right\} K_h(U_i - u_0).$$

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

$$\begin{aligned} \mathbf{H}^{-1}\, \ell_n''\left(\widetilde{\boldsymbol{\beta}}\right)\, \mathbf{H}^{-1} &= \mathbf{H}^{-1}\, \ell_n''(\boldsymbol{\beta})\, \mathbf{H}^{-1} + o_p(1) \\ &= \frac{1}{n}\sum_{i=1}^n q_2\left\{\widetilde{\mathbf{Z}}_i^T \boldsymbol{\beta},\, Y_i\right\}\, \mathbf{H}^{-1}\, \widetilde{\mathbf{Z}}_i\, \widetilde{\mathbf{Z}}_i^T\, \mathbf{H}^{-1}\, K_h(U_i - u_0) + o_p(1), \end{aligned} \tag{A.14}$$

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 $\mathbf{H}^{-1}\, \ell_n''(\boldsymbol{\beta})\, \mathbf{H}^{-1}$, we obtain

$$\mathbf{H}^{-1}\, \ell_n''\left(\widetilde{\boldsymbol{\beta}}\right)\, \mathbf{H}^{-1} \quad = \quad E\left[q_2\left\{\widetilde{\mathbf{Z}}^T \boldsymbol{\beta},\, Y\right\}\left(\begin{matrix} 1 \\ (U-u_0)/h \end{matrix}\right)^{\otimes 2} \otimes \mathbf{X}\, \mathbf{X}^T\, K_h(U-u_0)\right] + o_p(1)$$

$$
\begin{aligned}
&= E\left[q_2\left\{\tilde{\mathbf{Z}}^T\boldsymbol{\beta},\ m(U,\mathbf{X})\right\}\begin{pmatrix}1\\(U-u_0)/h\end{pmatrix}^{\otimes 2}\otimes\mathbf{X}\,\mathbf{X}^T K_h(U-u_0)\right]+o_p(1)\\
&= -\Delta+o_p(1), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(A.15)}
\end{aligned}
$$

where Δ is defined in (2.6). Recall that $\widehat{\boldsymbol{\beta}}_{\rm OS}=\widehat{\boldsymbol{\beta}}_0-\left\{\ell_n''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\ell_n'\left(\widehat{\boldsymbol{\beta}}_0\right)$ (see (2.2)). By the Taylor expansion, we have

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

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

$$
\begin{aligned}
\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_{\rm OS}-\boldsymbol{\beta}\right) &= \mathbf{H}\left(\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}\right)-\mathbf{H}\left\{\ell_n''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\mathbf{H}\,\mathbf{H}^{-1}\ell_n'\left(\widehat{\boldsymbol{\beta}}_0\right)\\
&= \left[\mathbf{I}-\mathbf{H}\left\{\ell_n''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\mathbf{H}\,\mathbf{H}^{-1}\ell_n''\left(\widetilde{\boldsymbol{\beta}}^*\right)\mathbf{H}^{-1}\right]\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}\right)\\
&\quad -\mathbf{H}\left\{\ell_n''\left(\widehat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\mathbf{H}\,\mathbf{H}^{-1}\ell_n'(\boldsymbol{\beta}). \qquad\qquad \text{(A.16)}
\end{aligned}
$$

Therefore, by (A.15) and (A.16), we have

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

which, in conjunction with (A.3), (A.9), (A.12) and (A.13), implies that

$$
\sqrt{n\,h}\,\mathbf{H}\left(\widehat{\boldsymbol{\beta}}_{\rm OS}-\boldsymbol{\beta}\right)=\Delta^{-1}\,W_n+o_p(1)=\widehat{\boldsymbol{\beta}}^*+o_p(1). \qquad \text{(A.17)}
$$

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

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," tentatively accepted by *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.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998), "Local estimating equations", *Journal of the American Statistical Association*, **93**, 214-227.

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. (2000), "Functional linear models for longitudinal data," *Journal of the Royal Statistical Society, Series B*, to appear.

Fan, J. and Zhang, W. (2000), "Statistical estimation in varying-coefficient models," *The Annals of Statistics*, to appear.

Fan, J., Zhang, C. and Zhang, J. (1999), "Sieve likelihood ratio statistics and Wilks phenomenon," *Technical Report*, Department of Statistics, UCLA.

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

Hong, Y. and Lee, T.-H. (1999), "Inference and forecast of exchange rates via generalized spectrum and nonlinear times series models," *Manuscript.*

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.

Kauermann, G. and Tutz, G. (1999), "On model diagnostics using varying coefficient models", *Biometrika*, **86**, 119-128.

Lee, T.-H. and Ullah, A. (1999), "Nonparametric bootstrap tests for neglected nonlinearity in time series regression models," *Manuscript.*

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

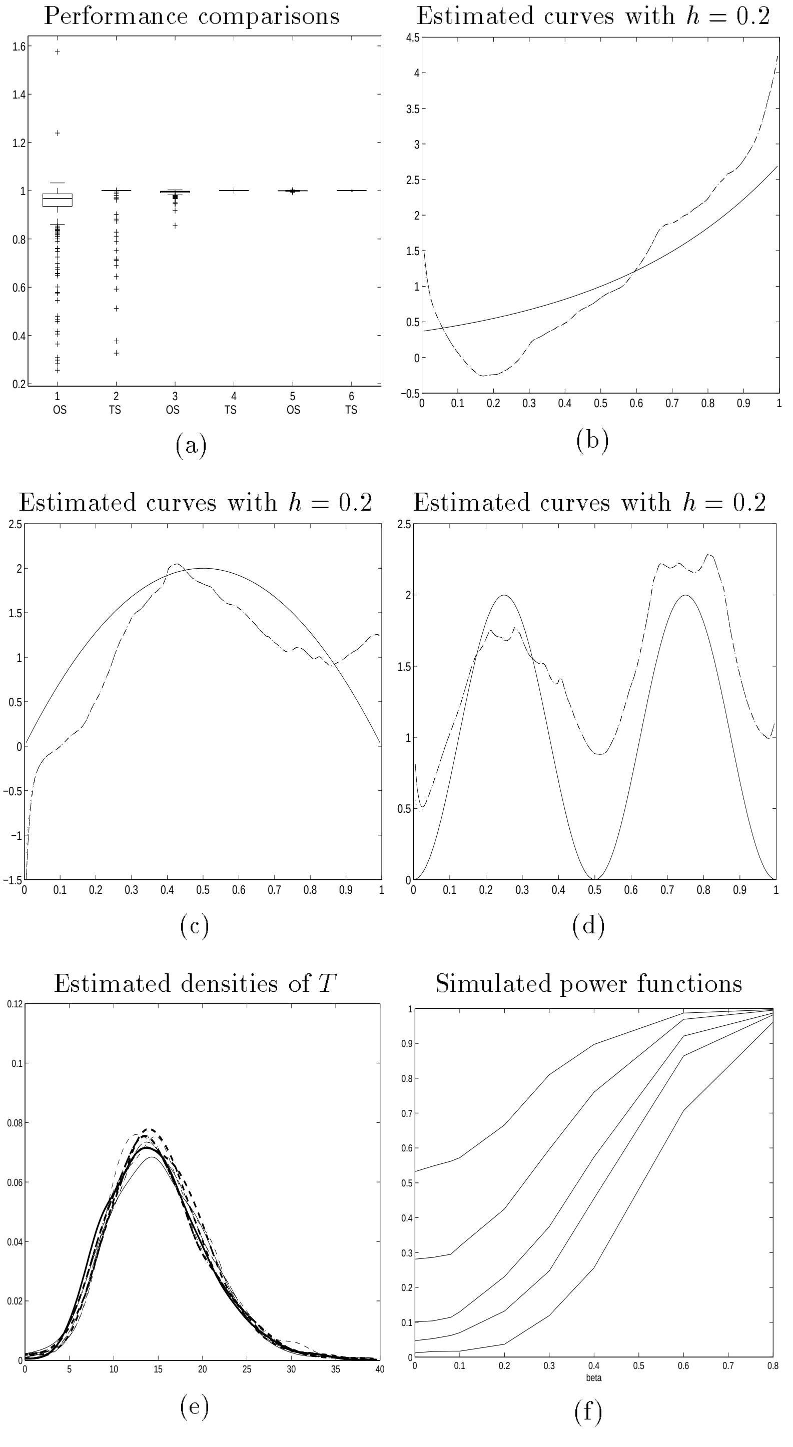
McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed, London: Chapman and Hall.

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.

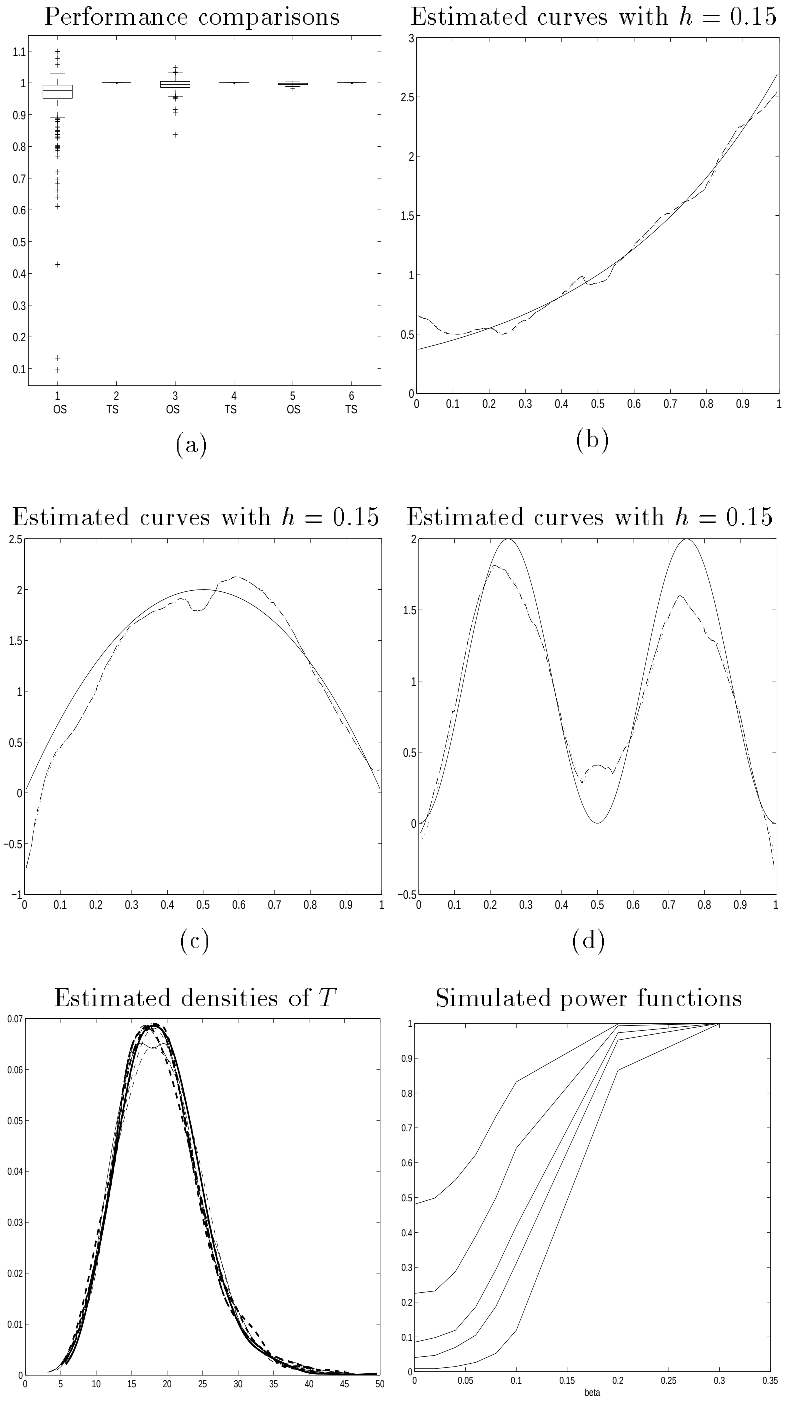


(a) (b)

(c) (d)

(e) (f)

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 curves) and for conditional null distributions (thin curves). (f) The power functions of the test statistic* T.



Performance comparisons

Estimated curves with $h = 0.15$

(a)

(b)

Estimated curves with $h = 0.15$

Estimated curves with $h = 0.15$

(c)

(d)

Estimated densities of T

Simulated power functions

(e)

(f)

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

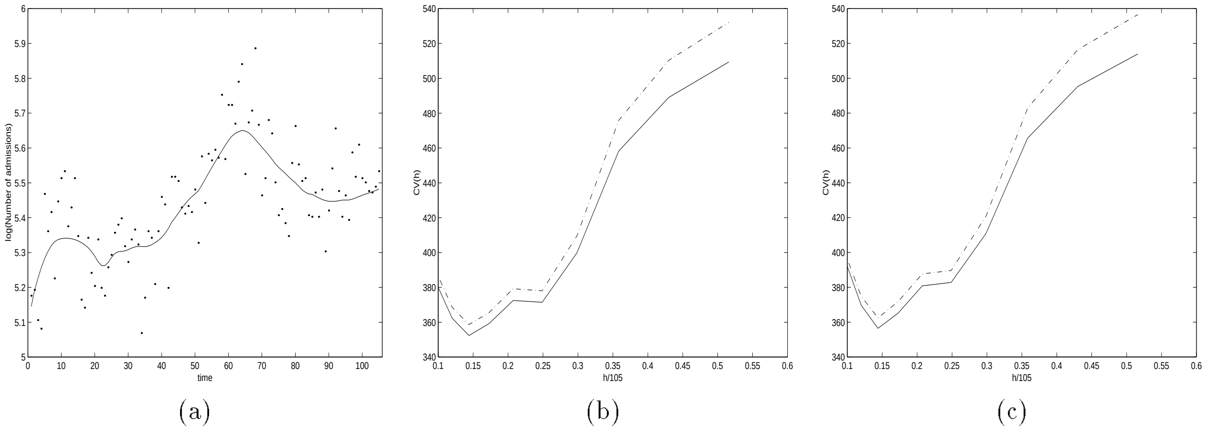


Figure 3: *(a) The scatterplot of log transformation of environmental data set studied in §3.3. The curve is the estimate of $a_1(t) + a_2(t)\,\bar{x}_1 + a_3(t)\,\bar{x}_2 + a_4(t)\,\bar{x}_3$, where $\bar{x}_j$ is the average pollutant level x_j. (b) The plot of the cross-validation functions $CV_1(h)$ (solid line) and $CV_2(h)$ (dashdot line) against bandwidth. (c) The same as those in (b), but the cross-validation is based on random partitions of the data set.*

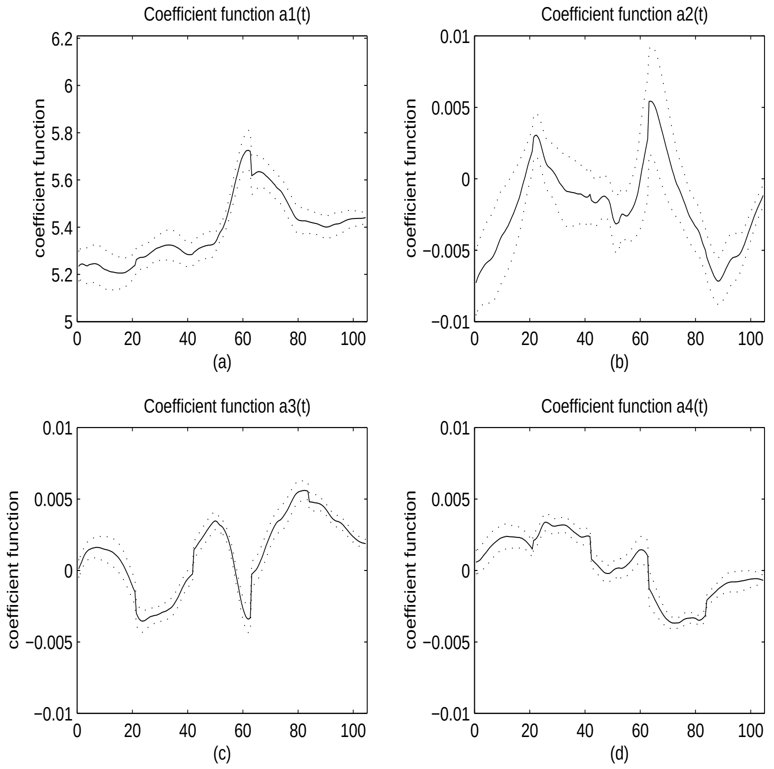
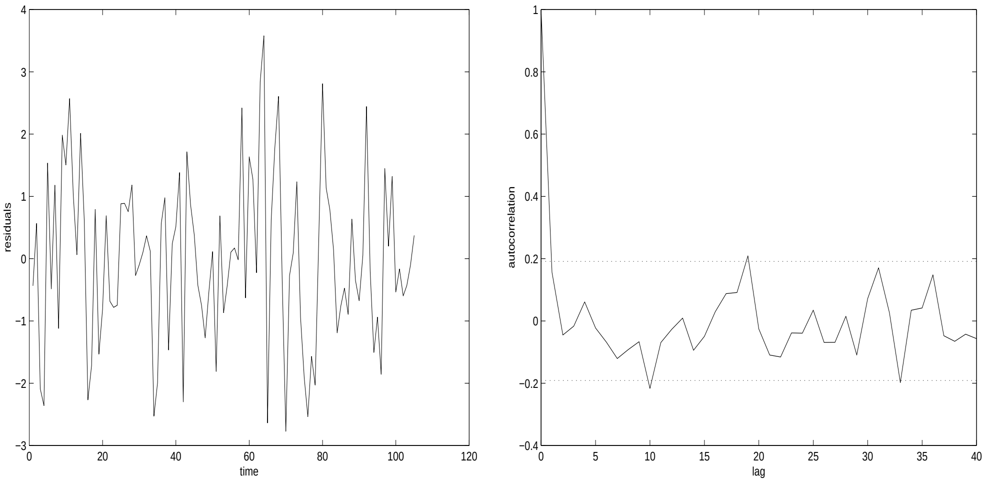


Figure 4: *The estimated coefficient functions via the one-step approach with bandwidth chosen by the CV. The dashed curves are the estimated function plus/minus twice estimated standard errors.*



(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 dashed curves are $\pm 1.96/\sqrt{n}$, where n is the sample size.*

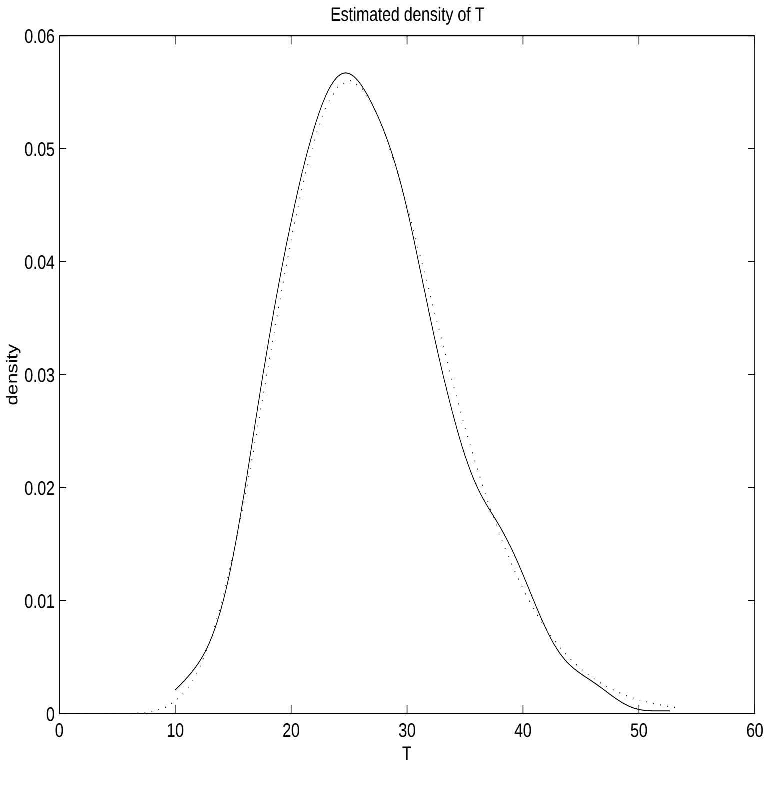


Figure 6: *The estimated density of T by Monte Carlo simulation. The solid curve is the estimated density, and the dashed curve stands for the density of chi-squared distribution with degrees of freedom 27.*

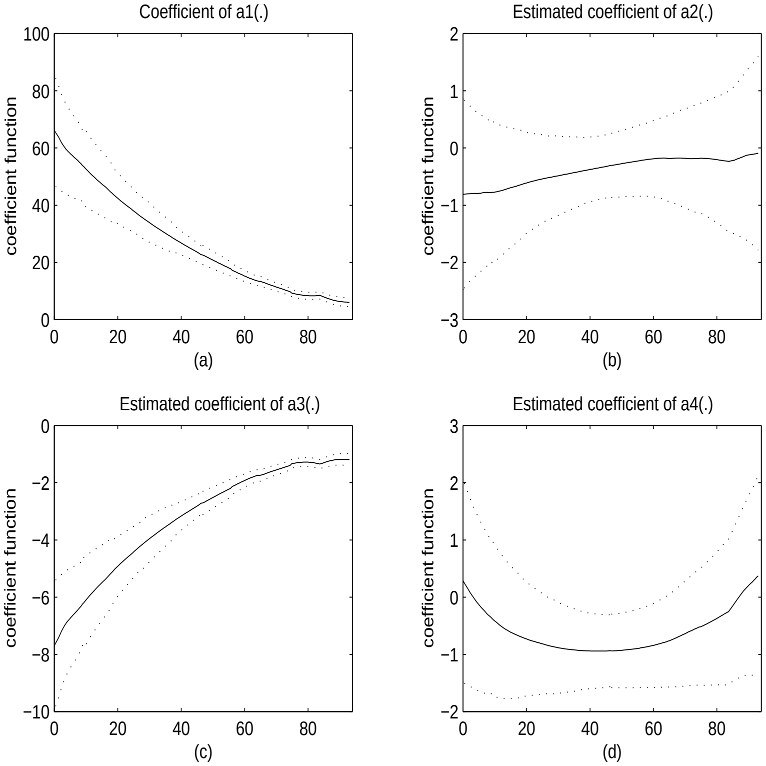


Figure 7: *The estimated coefficient functions (the solid curves) via one-step approach with bandwidth chosen by the CV. The dot curves are the estimated functions plus/minus twice estimated standard errors.*