Journal of Statistical Planning and Inference 139 (2009) 1056 – 1072

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Double-smoothing for bias reduction in local linear regression

Hua He*, Li-Shan Huang

Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14642, USA

ARTICLE INFO

Article history:
Received 24 October 2007
Received in revised form
23 May 2008
Accepted 19 June 2008
Available online 2 July 2008

Keywords:
Asymptotic bias
Asymptotic variance
Edge effect
Local polynomial regression
Mean square error
Nonparametric regression

ABSTRACT

Local linear regression involves fitting a straight line segment over a small region whose midpoint is the target point x, and the local linear estimate at x is the estimated intercept of that straight line segment, with an asymptotic bias of order h^2 and variance of order $(nh)^{-1}$ (h is the bandwidth). In this paper, we propose a new estimator, the double-smoothing local linear estimator, which is constructed by integrally combining all fitted values at x of local lines in its neighborhood with another round of smoothing. The proposed estimator attempts to make use of all information obtained from fitting local lines. Without changing the order of variance, the new estimator can reduce the bias to an order of h^4. The proposed estimator has better performance than local linear regression in situations with considerable bias effects; it also has less variability and more easily overcomes the sparse data problem than local cubic regression. At boundary points, the proposed estimator is comparable to local linear regression. Simulation studies are conducted and an ethanol example is used to compare the new approach with other competitive methods.

1. Introduction

Statistical smoothing techniques are often used to explore unknown trends. Frequently used nonparametric smoothing techniques include local polynomial regression (Fan and Gijbels, 1996; Wand and Jones, 1995), smoothing splines (Eubank, 1999), and penalized splines (Ruppert et al., 2003). These methods provide flexible modeling tools as there is no assumption on the functional form for the unknown regression function. Among these methods, local polynomial regression, especially local linear regression, enjoys excellent numerical as well as theoretical properties. Comparing local linear regression to high-order local polynomial regression, local linear regression not only is less likely to encounter sparse data problem because its design matrix is less likely to be singular or nearly singular, but also it is easier to implement corrections such as those based on ridging and shrinkage.

In this paper, we propose a new estimator which combines fitted values at a target point, say t, of all local lines in its neighborhood. The new estimator involves two steps of smoothing (weighted averaging), and we name it the double-smoothing local linear estimator. In the first step of smoothing, each local line is fitted by minimizing a weighted sum of squares as when fitting local linear regression, and the weight function controls how much weight is given to the observations. In the second step of smoothing, the new estimator is obtained by combining all fitted values at t of the first-step fitted local lines through a weighted integral. In the second step, another weight function, which does not need to be the same as in the first step, is used to control the weights given to those fitted values. In contrast to using only an intercept in local linear regression, the proposed estimator makes use of all information obtained from local lines in its neighborhood. As shown in Theorem 1 in Section 3, the new estimator has greater bias reduction than local linear regression; the asymptotic bias can be reduced from an order of h^2 to an order of h^4

* Corresponding author. Tel.: +1 585 4556378; fax: +1 585 2731031.
E-mail addresses: huahe@bst.rochester.edu, hua_he@urmc.rochester.edu (H. He), lhuang@bst.rochester.edu (L.-S. Huang).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.06.011

without changing the convergence rate of the asymptotic variance, where h is the bandwidth. Although the new estimator has the same convergence rate on the asymptotic bias as that for local cubic regression, it has less variability. Furthermore, since the design matrix is only used in the first step of smoothing, similar to local linear regression, the new estimator has the advantage of overcoming the sparse data problem. We note that the estimator proposed by Choi and Hall (1998) is a special case of the new estimator; we also note that when the same weight function for both steps of smoothing is applied, the estimator obtained by double-smoothing yields exactly the same estimated values for the mean function at the observed values of X as the `projected' response obtained by Huang and Chen (2008).

After a brief review of local linear regression and some related work in Section 2, in Section 3 we introduce the double-smoothing local linear estimator and develop its main properties inside the support of the predictor X. Since boundary effect is an important issue in smoothing techniques, we present a thorough discussion on the boundary estimation for the proposed estimator in Section 4. To illustrate the finite sample performance of the proposed estimator and compare the new estimator with local linear regression, local cubic regression and Choi and Hall's estimator, simulation studies are conducted in Section 5 and a real data example is presented in Section 6. Finally the paper concludes with a discussion in Section 7.

2. Local linear regression

Suppose we have independent observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from the model:

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where X and ε are independent and ε has mean 0 and variance 1. At a target point x, we want to estimate the regression mean $m(x) = E(Y|X = x)$. Local linear regression assumes that in a neighborhood of x, $m(X_i)$ can be approximated by $m(X_i) \approx m(x) + m'(x)(X_i - x)$. Then the local linear regression estimator (LL) is obtained by minimizing a weighted sum of squares:

$$\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1(X_i - x))\}^2 K\left(\frac{X_i - x}{h}\right), \tag{2}$$

where $K(\cdot)$ is a symmetric density function, i.e., $K \geqslant 0$, $\int K = 1$, and $K(-x) = K(x)$. The density function $K(\cdot)$ generally gives more weight to observations closer to x. The minimizing pair of (β_0, β_1), depending on the point x as well as on the data $\{(X_i, Y_i),\ i = 1, 2, \ldots, n\}$, is denoted as $(\widehat{\beta}_0(x), \widehat{\beta}_1(x))$. Local linear regression estimator, denoted by $\widetilde{m}(x)$, uses only $\widehat{\beta}_0(x)$ to estimate the value of $m(\cdot)$ at point x, i.e., $\widetilde{m}(x) = \widehat{\beta}_0(x)$.

If the weight function $K(\cdot)$ is supported on a compact interval, say $[-1, 1]$, then it is obvious that only observations in the region $[x - h, x + h]$ will be used to estimate $\widetilde{m}(x)$. Thus the bandwidth h is a smoothing parameter which controls the size of the neighborhood of local smoothing. If $[x - h, x + h]$ is included in the support of the design density, i.e., if the point x is an interior point, the asymptotic bias and variance for local linear regression (Fan and Gijbels, 1996) are given by

$$\begin{aligned} &\text{bias}\{\widetilde{m}(x)|X_1, \ldots, X_n\} = h^2 \frac{m''(x)}{2}\mu_2 + o_p(h^2), \\ &\text{var}\{\widetilde{m}(x)|X_1, \ldots, X_n\} = (nh)^{-1}\frac{v_0}{f(x)}\sigma^2(x) + o_p\{(nh)^{-1}\}, \end{aligned} \tag{3}$$

where $f(x)$ is the density of the covariate X, called the "design density", $\mu_j = \int u^j K(u)\,du$ and $v_j = \int u^j K^2(u)\,du$, $j = 0, 1, 2, \ldots$.

If $[x - h, x + h]$ is not entirely contained in the support of the design density, then (3) is no longer true. Such point x is called a boundary point. Most smoothing techniques have worse behavior at boundary points, and hence special handling is in general required. This is known as the boundary effect problem. There is an extensive literature on how to correct boundary effect; see, for example, Cline and Hart (1991), Cheng et al. (1997) and Cowling and Hall (1996). Although, as discussed in Fan and Gijbels (1996), local linear regression can adapt automatically to estimation at the boundary points, and hence no boundary modification is needed, the expression of the asymptotic bias and variance of $\widetilde{m}(x)$ in the boundary region is different.

Assume that the design density has a bounded support $[0, 1]$, and $K(\cdot)$ is supported on $[-1, 1]$. A left boundary point has the form $x = ch$ with $0 \leqslant c < 1$, whereas a right boundary point is of the form $x = 1 - ch$. At a left boundary point $x = ch$ (the right boundary point would be similar), the asymptotic bias and variance for local linear regression (Fan and Gijbels, 1996) are given by

$$\begin{aligned} &\text{bias}\{\widetilde{m}(x)|X_1, \ldots, X_n\} = h^2 \frac{m''(x)}{2}B_0(c) + o_p(h^2), \\ &\text{var}\{\widetilde{m}(x)|X_1, \ldots, X_n\} = (nh)^{-1}\frac{V_0(c)}{f(x)}\sigma^2(x) + o_p\{(nh)^{-1}\}, \end{aligned} \tag{4}$$

where $B_0(c) = (\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c})/(\mu_{0,c}\mu_{2,c} - \mu_{1,c}^2)$ and $V_0(c) = \int_{-c}^{1}(\mu_{2,c} - u\mu_{1,c}^2)K^2(u)\,du/(\mu_{2,c}\mu_{0,c} - \mu_{1,c}^2)^2$ with $\mu_{j,c} = \int_{-c}^{1} u^j K(u)\,du$.