Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 2808–2828
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Robust bandwidth selection in semiparametric partly linear regression models: Monte Carlo study and influential analysis

Graciela Boente*, Daniela Rodriguez

Departamento de Matemáticas, Instituto de Cálculo, FCEyN, Universidad de Buenos Aires and CONICET, Argentina

Received 28 July 2006; received in revised form 15 October 2007; accepted 15 October 2007
Available online 26 October 2007

Abstract

In this paper, under a semiparametric partly linear regression model with fixed design, we introduce a family of robust procedures to select the bandwidth parameter. The robust plug-in proposal is based on nonparametric robust estimates of the vth derivatives and under mild conditions, it converges to the optimal bandwidth. A robust cross-validation bandwidth is also considered and the performance of the different proposals is compared through a Monte Carlo study. We define an empirical influence measure for data-driven bandwidth selectors and, through it, we study the sensitivity of the data-driven bandwidth selectors. It appears that the robust selector compares favorably to its classical competitor, despite the need to select a pilot bandwidth when considering plug-in bandwidths. Moreover, the plug-in procedure seems to be less sensitive than the cross-validation in particular, when introducing several outliers. When combined with the three-step procedure proposed by Bianco and Boente [2004. Robust estimators in semiparametric partly linear regression models. J. Statist. Plann. Inference 122, 229–252] the robust selectors lead to robust data-driven estimates of both the regression function and the regression parameter.

Keywords: Asymptotic properties; Bandwidth selectors; Kernel weights; Partly linear models; Robust estimation; Smoothing techniques

1. Introduction

Partly linear models have become an important tool when modelling biometric data, since they combine the flexibility of nonparametric models and the simple interpretation of the linear ones. These models assume that we have a response $y_i \in \mathbb{R}$ and covariates or design points $(\mathbf{x}_i^{\mathrm{T}}, t_i)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ satisfying

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad 1 \leqslant i \leqslant n, \tag{1}$$

with the errors ε_i independent and independent of $(\mathbf{x}_i^{\mathrm{T}}, t_i)^{\mathrm{T}}$. The semiparametric nature of model (1) offers more flexibility than the standard linear model, when modelling a complicated relationship between the response variable with one of the covariates. At the same time, they keep a simple functional form with the other covariates avoiding the "curse of dimensionality" existing in nonparametric regression.

* Corresponding author. Tel./fax: +54 11 45763375.
E-mail address: gboente@dm.uba.ar (G. Boente).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.10.017

In many situations, it seems reasonable to suppose that a relationship between the covariates $\mathbf{x}$ and t exists, so as in Speckman (1988), Linton (1995) and Aneiros-Pérez and Quintela del Río (2002), we will assume that for $1 \leqslant j \leqslant p$

$$x_{ij} = \phi_j(t_i) + \eta_{ij}, \quad 1 \leqslant i \leqslant n, \tag{2}$$

where the errors η_{ij} are independent. Moreover, the design points t_i will be assumed to be fixed.

Several authors have considered the semiparametric model (1). See, for instance, Denby (1986), Rice (1986), Robinson (1988), Speckman (1988) and Härdle et al. (2000) among others.

All these estimators, as most nonparametric estimators, depend on a smoothing parameter that should be chosen by the practitioner. As it is well known, large bandwidths produce estimators with small variance but high bias, while small values produce more wiggly curves. This trade-off between bias and variance lead to several proposals to select the smoothing parameter, such as cross-validation procedures and plug-in methods. Linton (1995), using local polynomial regression estimators, obtained an asymptotic expression for the optimal bandwidth in the sense that it minimizes a second order approximation of the mean square error of the least squares estimate, $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}(h)$, of $\boldsymbol{\beta}$. This expression depends on the regression function we are estimating and on parameters which are unknown, such as the standard deviation of the errors. More precisely, for any $\mathbf{c} \in \mathbb{R}^p$, let $\sigma^2 = \sigma_\varepsilon^2 \mathbf{c}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \mathbf{c}$ be the asymptotic variance of $U = \mathbf{c}^{\mathrm{T}} n^{1/2} (\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}(h) - \boldsymbol{\beta})$, and $nMSE(h) = EU^2/\sigma^2$ its standardized mean square error. For the sake of simplicity, assume that the smoothing procedure corresponds to local means and that the design points are almost uniform design points, i.e., $\{t_i\}_{i=1}^n$ are fixed design points in $[0, 1]$, $0 \leqslant t_1 \leqslant \cdots \leqslant t_n \leqslant 1$, such that $t_0 = 0$, $t_{n+1} = 1$ and $\max_{1 \leqslant i \leqslant n+1} |(t_i - t_{i-1}) - 1/n| = \mathrm{O}(n^{-\delta})$ for some $\delta > 1$. Then, under general conditions, we have that, for $v \geqslant 2$,

$$MSE(h) = n^{-1}\{1 + (nh)^{-1} A_2 + \mathrm{o}(n^{-2\mu}) + (n^{1/2} h^{2v} A_1 + \mathrm{o}(n^{-\mu}))^2\},$$

where $\mu = (4v - 1)/(2(4v + 1))$, $\boldsymbol{\phi}^{(v)}(t) = (\phi_1^{(v)}(t), \ldots, \phi_p^{(v)}(t))^{\mathrm{T}}$, $\alpha_v(K) = \int u^v K(u)\, \mathrm{d}u$, $K_*(u) = K * K(u) - 2K(u)$ and

$$A_1 = \alpha_v^2(K)(v!)^{-2} \sigma^{-1} \mathbf{c}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \int_0^1 g^{(v)}(t) \boldsymbol{\phi}^{(v)}(t)\, \mathrm{d}t, \quad A_2 = \int K_*^2(u)\, \mathrm{d}u.$$

Therefore, the optimal bandwidth in the sense of minimizing the asymptotic $MSE(h)$, is given by $h_{\mathrm{opt}} = A_0 n^{-\pi}$, with $\pi = 2/(4v + 1)$ and $A_0 = (A_2/(4vA_1^2))^{\pi/2}$, i.e.,

$$A_0 = \left\{ \int K_*^2(u)\, \mathrm{d}u \Big/ \left[4v \left(\sigma^{-1} \mathbf{c}^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1} \alpha_v^2(K)(v!)^{-2} \int_0^1 g^{(v)}(t) \boldsymbol{\phi}^{(v)}(t)\, \mathrm{d}t \right)^2 \right] \right\}^{\pi/2}. \tag{3}$$

Linton (1995) considered a plug-in approach to estimate the optimal bandwidth and showed that it converges to the optimal one, while Aneiros-Pérez and Quintela del Río (2002) studied the case of dependent errors.

It is well known that, both in linear regression and in nonparametric regression, least squares estimators can be seriously affected by anomalous data. The same statement holds for partly linear models, where large values of the response variable y_i can cause a peak on the estimates of the smooth function g in the neighborhood of t_i. Moreover, large values of the response variable y_i combined with high leverage points $\mathbf{x}_i$ produce also, as in linear regression, breakdown of the classical estimates of the regression parameter $\boldsymbol{\beta}$. To overcome that problem, Bianco and Boente (2004) considered a three-step robust estimate for the regression parameter and the regression function. Besides, for the nonparametric regression setting, i.e., when $\boldsymbol{\beta} = 0$, the sensitivity of the classical bandwidth selectors to anomalous data was discussed by several authors, such as, Leung et al. (1993), Wang and Scott (1994), Boente et al. (1997), Cantoni and Ronchetti (2001) and Leung (2005).

In this paper, we consider a robust plug-in selector for the bandwidth, under the partly linear model (1) which converges to the optimal one and leads to robust data-driven estimates of the regression function g and the regression parameter $\boldsymbol{\beta}$. We derive an expression analoguous to (3) for the optimal bandwidth of the three-step estimator introduced in Bianco and Boente (2004). As for its linear relative, this expression will depend on the derivatives of the functions g and $\boldsymbol{\phi}$. In Section 2, we review some of the proposals given to estimate robustly the derivatives of the regression function under a nonparametric regression model. The robust plug-in bandwidth selector for the partial linear model is introduced in Section 3 together with a robust cross-validation procedure. In Section 4, for small samples, the behavior