Journal of Statistical Planning and Inference 146 (2014) 102–112

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Statistical inference in the partial linear models with the double smoothing local linear regression method

Hua He [a], Wan Tang [a], Guoxin Zuo [a,b,*]

[a] Department of Biostatistics and Computational Biology, University of Rochester, USA
[b] School of Mathematics and Statistics, Central China Normal University, China

ARTICLE INFO

Article history:
Received 18 August 2011
Received in revised form
6 August 2013
Accepted 9 October 2013
Available online 18 October 2013

Keywords:
Difference-based estimation
Double smoothing
Local linear regression
Partial linear model

ABSTRACT

Statistical inference about a partial linear model includes two parts, one linear and the other nonparametric. The double smoothing method proposed by He and Huang (2009) is a progressive local smoothing method for nonparametric curve estimation. In this paper, we discuss its extension to partial linear models, accompanied with difference-based estimation method for the linear part. Asymptotic theory of the proposed method is developed. The results of simulation studies and real data examples demonstrate that our approach is effective even for data with moderate sample sizes.

1. Introduction

As a very useful compromising tool between nonparametric and parametric regression, semiparametric regression is getting more popular recently. In a semiparametric regression model there are two components; one is parametric and the other is nonparametric. Both the modeling flexibility and the simplicity in implementation make these models very attractive. One of the most popular models, partial linear model, plays an important role in various applications (Engle et al., 1986; Green et al., 1985). The linear form assumed in the parametric part makes it simple to interpret while the flexibility of the model still holds by the nonparametric part. Much statistical literature focused on the inference about the parametric part for partial linear models, treating the nonparametric part as an infinite-dimensional nuisance parameter (Chen, 1988; Bhattacharya and Zhao, 1997; Rice, 1986; Heckman, 1986; Robinson, 1988). However, the nonparametric part also plays an important role in data analysis, such as the relationship between the logarithm of plasma beta-carotene level and age of person in nutriology (see the second example in Section 5 below for details). Appropriate estimations of the nonparametric part can reveal the internal relationship between the response and explanatory variables (Marra and Radice, 2010).

In most of the inference procedures of a partial linear model, the parametric part and nonparametric part are jointly considered and various nonparametric methods are applied. The penalized least squares methods were considered by Engle et al. (1986), Heckman (1986), and Wahba (1984). Following the kernel type method (Robinson, 1988; Speckman, 1988), other smoothing methods, such as the local polynomial method (Hamilton and Truong, 1997; Aneiros-Pérez and Vilar-Fernández, 2008), and the wavelet approach (Chang and Qu, 2004) are also applied to partial linear models. Efficiency and consistency of these estimations for the partial linear model were systematically discussed by Härdle et al. (2000).

* Corresponding author at: School of Mathematics and Statistics, Central China Normal University, China. Tel.: +86 13 971655959; fax: +86 27 67867452.
E-mail addresses: zuoguoxin@gmail.com, zuogx@mail.ccnu.edu.cn (G. Zuo).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.10.004

Yatchew (1997) applied the simple difference-based method to estimate the parametric component of partial linear models. Under some mild conditions on the nonparametric smoothing function, the least squares estimation for the parametric part was obtained directly. Wang et al. (2011) further discussed the efficiency of the difference-based estimation method for partial linear models. Zhao and You (2011) discussed difference based estimation for the partially linear regression model with measurement errors. The estimation of the parametric part provides us another approach to estimate the nonparametric part using available nonparametric methods.

Recently, He and Huang (2009) suggested a double smoothing local linear (DSLL) method in univariate nonparametric regressions to reduce the bias of the estimation. In this paper, we extend the DSLL method to partial linear models, and show that, asymptotically, the DSLL estimates for partial linear models have similar advantages over their local linear (LL) counterpart as in the univariate nonparametric regression cases. Furthermore, the estimations of the parametric part of the partial linear models have little effect on the performance of the DSLL estimations of the nonparametric part; the DSLL method based on the $\sqrt{n}-$consistent difference-based estimate is asymptotically equivalent to the DSLL estimate when the parameters of the linear part are known. The above asymptotic results are also supported by simulation studies under small to moderate sample sizes. In the rest of the paper, the difference-based estimation method for partial linear models is briefly reviewed and the DSLL method is generalized to partial linear models in Section 2. The main result about the DSLL method for partial linear regression models is presented in Section 3. Simulation studies are carried out in Section 4 and the methods are applied to two real data sets in Section 5. In Section 6, we discuss some inference problems for further study.

2. Statistical inference of the partial linear model

2.1. Partial linear model

Let Y be the response variable of interest, and (U,X) the explanatory variable vectors such that U is a scalar and X a p-dimensional vector. A partial linear model has the form

$$Y = \alpha(U) + X^T\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\alpha(\cdot)$ is an unknown nonparametric smooth function, and ε, independent with U and X, is a random error with mean zero and finite variance σ_ε^2. Thus, a partial linear model can be thought as an extension of the multiple linear regression model. The nonparametric part $\alpha(U)$ provides the flexibility to accommodate the nonlinear nature. A two-step procedure can be applied for inference; the parameter $\boldsymbol{\beta}$ is first estimated via the difference-based method and then $\alpha(\cdot)$ is estimated using smoothing methods. For the convenience of the readers we briefly review the difference-based method (for details see Yatchew, 1997; Wang et al., 2011).

2.2. Difference-based estimate of $\boldsymbol{\beta}$

Yatchew (1997) applied the difference-based method to the partial linear model and showed that the estimator of parameter is $\sqrt{n}-$consistent if the dimension of covariate in the unknown smoothing function is at most 3.

Let $(Y_i, U_i, X_i^T)_{i=1,\ldots,n}$ be an i.i.d. sample, and $(Y_i^*, U_i^*, X_i^{*T})_{i=1,\ldots,n}$, the ordered sample based on U_i, i.e., $U_1^* \le \cdots \le U_n^*$. Let $d_0, \ldots, d_m (m \ge 1)$ be a sequence of difference coefficients satisfying

$$\sum_{j=0}^{m} d_j = 0, \qquad \sum_{j=0}^{m} d_j^2 = 1. \tag{2}$$

The corresponding difference sequences in X and Y are defined as

$$\Delta\mathbf{X} = \left(\sum_{j=0}^{m} d_j X_{t+m-j}^{*T}\right)_{1\le t\le n-m} \quad \text{and} \quad \Delta\mathbf{Y} = \left(\sum_{j=0}^{m} d_j Y_{t+m-j}^{*}\right)_{1\le t\le n-m},$$

respectively. The basic idea of the difference-based method is that $\Delta\mathbf{Y} = \Delta\mathbf{X}^T\boldsymbol{\beta} + \Delta\alpha(U) + \Delta\varepsilon$, where $\Delta\alpha(U)$ and $\Delta\varepsilon$ are defined similar to $\Delta\mathbf{Y}$ and $\Delta\mathbf{X}$. The difference in $\alpha(U)$, $\Delta\alpha(U)$, can be integrated with the difference error $\Delta\varepsilon$, and $\boldsymbol{\beta}$ can be estimated by the regression of $\Delta\mathbf{Y}$ on $\Delta\mathbf{X}$. Therefore, the difference-based estimation of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_m = (\Delta\mathbf{X}^T\Delta\mathbf{X})^{-1}\Delta\mathbf{X}^T\Delta\mathbf{Y}.$$

The estimate $\widehat{\boldsymbol{\beta}}_m$ is asymptotic normal and $\sqrt{n}-$consistent, and asymptotic efficiency can be achieved under some mild regularity assumptions. More precisely, let $g(u,x)$ be the joint probability density function of U and X and $f(u)$ be the probability density function of U. Without loss of generality, we may assume that U has a compact support $D_U \subset R^1$. For example, we can always simply truncate U on the region of interest. We need the following assumption, similar to that of Yatchew (1997).

Assumption 1.

(a) The first derivative of the nonparametric function $\alpha(\cdot)$ is bounded on D_U.
(b) $f(u)$ is bounded away from 0 on its compact support D_U.
(c) The optimal difference $d_1, \ldots, d_m$ are selected as in Hall et al. (1990).
(d) X has a nondegenerate conditional distribution given U and $\Sigma_X^* = E(Cov(X|U))$ exists.