Local polynomial estimation in partial linear regression models under dependence

G. Aneiros-Pérez*, J.M. Vilar-Fernández

Departamento de Matemáticas, Universidade da Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

Abstract

A regression model whose regression function is the sum of a linear and a nonparametric component is presented. The design is random and the response and explanatory variables satisfy mixing conditions. A new local polynomial type estimator for the nonparametric component of the model is proposed and its asymptotic normality is obtained. Specifically, this estimator works on a prewhitening transformation of the dependent variable, and the results show that it is asymptotically more efficient than the conventional estimator (which works on the original dependent variable) when the errors of the model are autocorrelated. A simulation study and an application to a real data set give promising results.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Asymptotic normality; Mixing processes; Nonparametric regression; Partial linear regression

1. Introduction

A special class of semiparametric regression models which are flexible, overcome (or reduce) the "curse of dimensionality" and allow easy interpretation of the effect of each explanatory variable on the response variable was proposed by Engle et al. (1986). This class of models, known as partial linear regression (PLR) models, assumes that the regression function is the sum of a linear and a nonparametric component, that is,

$$Y_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + m(\mathbf{T}_i) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id_0})^{\mathrm{T}}$ and $\mathbf{T}_i = (T_{i1}, \ldots, T_{id_1})^{\mathrm{T}}$ ($d_0 \geqslant 1$ and $d_1 \geqslant 1$) are vectors of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_0})^{\mathrm{T}}$ is a vector of unknown real parameters, m is an unknown smooth real function and $\{\varepsilon_i\}$ are the random errors satisfying

$$E(\varepsilon_i|\mathbf{X}_i, \mathbf{T}_i) = 0 \quad (i = 1, \ldots, n). \tag{2}$$

The PLR model has been studied extensively for i.i.d. data (see, for example, Heckman, 1986; Rice, 1986, for spline type estimators; Speckman, 1988; Robinson, 1988, for kernel type estimators; Linton, 1995, for local polynomial type estimators; Bianco and Boente, 2004, for robust estimation) as well as for dependent data (see, for example, Engle et al., 1986, for spline type estimators; Gao, 1995; Aneiros-Pérez et al., 2004, for kernel type estimators). In general,

---

* Corresponding author. Tel.: +34 981167000x1229; fax: +34 981167160.
  E-mail address: ganeiros@udc.es (G. Aneiros-Pérez).

these works propose different estimators for $\boldsymbol{\beta}$ and/or m in (1) and study their consistency and asymptotic normality. As occurs in nonparametric or semiparametric estimation, these estimators depend on smoothing parameters or bandwidths which must be selected from the observations. Therefore, automatic procedures to choose these bandwidths are needed. In spite of the extensive literature on PLR models, few works exist on the crucial problem of bandwidth selection. Some papers dealing with this topic are those of Linton (1995) and Aneiros-Pérez (2002) (both papers propose plug-in type selectors, the first paper focusing on i.i.d. data and the second on dependent data) and that of Aneiros-Pérez and Quintela-del-Río (2001a) (based on cross-validation ideas and focusing on dependent data). See also the work of Liang (2006) for some ideas on the strategy of bandwidth selection. In addition, PLR models have demonstrated their usefulness in many fields of applied sciences, such as economics, environmental studies, medicine, ... (see Härdle et al., 2000, for a monograph and applications of the PLR model).

This paper deals with the estimation of the nonparametric component m of the PLR model (1), assuming both random design on $\{(\mathbf{X}_i, \mathbf{T}_i)\}$ and mixing conditions on $\{(Y_i, \mathbf{X}_i, \mathbf{T}_i)\}$. For pure nonparametric regression models with random design (that is, model (1) where $\boldsymbol{\beta}$ is known and $\mathbf{T}_i$ are random vectors), Masry (1996a) and Xiao et al. (2003), among others, reported their findings. Masry obtained the asymptotic normality of the (conventional) local polynomial estimator, while Xiao et al. obtained that of a local polynomial estimator applied to a prewhitening transformation of the dependent variable, this transformation being estimated from the data. In conclusion, the estimator proposed by Xiao et al. takes into account the correlation structure of the error process and is asymptotically more efficient than that studied by Masry. Given this result, we will estimate m in the model (1) following the procedure proposed by Xiao et al., but applied to PLR models instead of to pure nonparametric models. As we will show, the asymptotic result proven by Xiao et al. holds for our estimator.

The construction of the estimator is presented in Section 2, while the asymptotic results and the conditions used to obtain these are given in Section 3. In Section 4, the finite behavior of the estimator is illustrated using a simulation study, and an application to a real data set is shown in Section 5 (all computations were performed with R). Concluding remarks appear in Section 6. Finally, Section 7 presents the proofs of our theorems while the Appendix gives some lemmas used in those proofs.

2. The estimator

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, and for any $(n \times s)$-matrix $\mathbf{A}$ $(s \geqslant 1)$ and real number $h > 0$, $\widetilde{\mathbf{A}}_h = (\mathbf{I} - \mathbf{W}_h)\mathbf{A}$, where $\mathbf{W}_h = (w_{h,j}(\mathbf{T}_i, \mathbf{T}_1, \ldots, \mathbf{T}_n))_{i,j}$ is an $n \times n$ matrix with $w_{h,j}(\cdot, \mathbf{T}_1, \ldots, \mathbf{T}_n)$ being a weight function that can take different forms. Throughout this paper, we will denote $w_{h,j}(\cdot) \equiv w_{h,j}(\cdot, \mathbf{T}_1, \ldots, \mathbf{T}_n)$.

From (1) and (2), we have that

$$Y_i - E(Y_i|\mathbf{T}_i) = (\mathbf{X}_i - E(\mathbf{X}_i|\mathbf{T}_i))^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \ldots, n). \tag{3}$$

Assuming that the conditional moments included in (3) are smooth functions of $\mathbf{T}_i$, these can be estimated using nonparametric estimators. Hence, an estimator of $\boldsymbol{\beta}$ can be obtained applying ordinary least squares (OLS) estimation to the model (3) after estimating those moments. This estimation gives

$$\widehat{\boldsymbol{\beta}}_{h_0} = (\widetilde{\mathbf{X}}_{h_0}^{\mathrm{T}} \widetilde{\mathbf{X}}_{h_0})^{-1} \widetilde{\mathbf{X}}_{h_0}^{\mathrm{T}} \widetilde{\mathbf{Y}}_{h_0}. \tag{4}$$

Finally, a nonparametric estimator for m in the PLR model (1) can be obtained by smoothing the points $\{(Y_j - \mathbf{X}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{h_0}, \mathbf{T}_j)\} \subset \mathbb{R} \times \mathbb{R}^{d_1}$, this estimator taking the expression

$$\widehat{m}_{h_0,h_1}(\mathbf{t}) = \sum_{j=1}^{n} w_{h_1,j}(\mathbf{t})(Y_j - \mathbf{X}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{h_0}). \tag{5}$$

(In (4) and (5), $h_0 > 0$ and $h_1 > 0$ are smoothing parameters or bandwidths that typically appear in nonparametric or semiparametric estimations.) The estimators (4) and (5) were proposed by Speckman (1988) and Robinson (1988) for i.i.d. data. Under suitable conditions, their asymptotic results (normality and rates of convergence) hold if the weights take the form

$$w_{h,j}(\mathbf{t}) = \frac{K((\mathbf{t} - \mathbf{T}_j)/h)}{\sum_{i=1}^{n} K((\mathbf{t} - \mathbf{T}_i)/h)}, \tag{6}$$