C. R. Acad. Sci. Paris, Ser. I 349 (2011) 1105–1109

Contents lists available at SciVerse ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

Asymptotic results for the linear parameter estimate in partially linear additive regression model

Résultats asymptotiques pour l'estimateur du paramètre linéaire dans le modèle de régression additif partiellement linéaire

Khalid Chokri [a,b], Djamal Louani [a,b]

[a] *L.S.T.A., Université de Paris 6, 4, place de Jussieu, 75252 Paris cedex 05, France*
[b] *L.S.T.A., Université de Reims, BP 1039, 51687 Reims cedex 2, France*

ARTICLE INFO

Article history:
Received 23 June 2011
Accepted after revision 14 September 2011
Available online 2 October 2011

Presented by Paul Deheuvels

ABSTRACT

In this Note, we study the linear part of the semi-parametric regression model defined by $Y_i = \mathbf{Z}_i^\top \beta + \sum_{j=1}^d m_j(X_{ij}) + \varepsilon_i$, $1 \leqslant i \leqslant n$, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top$ are vectors of explanatory variables, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown parameters, $m_1, \ldots, m_d$ are unknown univariate real functions, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random modelling errors with mean zero and finite variances. Using the nonparametric kernel technique combined with the marginal integration method to estimate the functions $(m_j)_{1\leqslant j\leqslant d}$ and the least-square error criterion to estimate the parameter β, we establish the asymptotic normality together with the iterated logarithm law of the estimate $\widehat{\beta}$ of β.

RÉSUMÉ

Cette Note est consacrée à l'étude de la partie linéaire du modèle de la régression partiellement linéaire défini par $Y_i = \mathbf{Z}_i^\top \beta + \sum_{j=1}^d m_j(X_{ij}) + \varepsilon_i$, $1 \leqslant i \leqslant n$, où $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top$ sont des vecteurs de variables explicatives, $\beta = (\beta_1, \ldots, \beta_p)^\top$ est un vecteur de paramètres inconnus, $m_1, \ldots, m_d$ sont des fonctions réelles univariées inconnues, et $\varepsilon_1, \ldots, \varepsilon_n$ sont les erreurs de modélisation supposées indépendantes de moyennes nulles et de variances finies. En utilisant la méthode du noyau accompagnée de la méthode d'intégration marginale pour estimer les fonctions $(m_j)_{1\leqslant j\leqslant d}$ et le critère des moindres carrés pour estimer le paramètre β, nous établissons la normalité asymptotique et la loi du logarithme itéré pour l'estimateur $\widehat{\beta}$ de β.

Version française abrégée

Soit $(\mathbf{X}_i, Y_i, \mathbf{Z}_i)_{i\geqslant 1}$ une suite de répliques indépendantes d'un vecteur aléatoire $(\mathbf{X}, Y, \mathbf{Z})$ à valeurs dans $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p$. Notons $\mathbf{g}$ la densité marginale de la composante $\mathbf{X}$. Les modèles de régression paramétrique fournissent de puissants outils pour la modélisation des données lorsque celles-ci s'y prêtent bien. Cependant, ces modèles peuvent être la source d'importants biais lorsqu'ils ne sont pas adéquats. Pour éliminer ces biais de modélisation, des méthodes non paramétriques ont été introduites permettant aux données elles mêmes de construire le modèle. Ces méthodes présentent, dans le cas multivarié,

E-mail addresses: khalid.chokri@etu.upmc.fr (K. Chokri), djamal.louani@upmc.fr (D. Louani).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2011.09.010

un handicap connu sous l'appellation de fléau de la dimension où la vitesse de convergence des estimateurs est une fonction décroissante de la dimension des covariables. Nous renvoyons à Stone [11] pour le détail lié à cette problématique.

Les modèles partiellement linéaires permettent d'allier les techniques paramétriques avec les méthodes non paramétriques pour la modélisation des données comportant une partie linéaire combinée à une partie non linéaire et se présentant sous la forme

$$Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon,$$

où $\beta \in \mathbb{R}^p$ est un paramètre vectoriel inconnu, m désigne la partie non linéaire du modèle et ε l'erreur de modélisation. Ici $\mathbf{Z}^\top$ indique le transposé du vecteur $\mathbf{Z}$. L'estimation par des méthodes non paramétriques de la partie non linéaire fait que, pour ce type de modèle, on est aussi confronté au fléau de la dimension.

Pour pallier le problème posé par le fléau de la dimension de ce modèle, on considère une structure additive de la fonction de régression m et on obtient un nouveau modèle sous la forme

$$Y = \mathbf{Z}^\top \beta + \sum_{l=j}^{d} m_j(X_j) + \varepsilon,$$

où X_j est le $j^{\text{ième}}$ composante du vecteur $\mathbf{X}$ et m_j est une fonction réelle univariée. Les quantités inconnues du modèle, qui doivent être estimées, sont le paramètre vectoriel β et les fonctions univariées $m_j, 1 \leqslant j \leqslant d$. Alors que le paramètre β est estimé en utilisant le critère quadratique moyen usuel, plusieurs méthodes ont été proposées dans la littérature pour estimer les composantes additives du modèle. Dans la suite de ce travail, nous utilisons la méthode de l'intégration marginale. Les estimateurs des quantités inconnues, qui en résultent, sont donnés plus bas dans les assertions (9), (10) et (11).

L'objet de cette Note, après la construction des estimateurs, est d'étudier les propriétés asymptotiques de normalité et de loi du logaritme itéré relatives à l'estimation du paramètre linéaire β. Les résultats obtenus sont exposés dans les Théorèmes 2.1 et 2.2.

Les modèles partiellement linéaires trouvent leurs applications dans de nombreux domaines alliant entre autres l'économie, la biologie, l'éducation et les sciences sociales. La littérature s'est enrichi de nombreux travaux proposant diverses études allant de diverses extensions de ce type de modèles à l'estimation des paramètres et aux tests d'adéquation. Notons aussi que des applications de ces modèles à des données réelles ont été réalisées et exposées dans de nombreux articles. Nous renvoyons, par exemple, aux travaux Robinson [10] et Härdle, Liang et Gao [8], et Aneiros-Pérez et Vieu [2] pour quelques études de cas.

1. Introduction

Let $(\mathbf{X}_i, Y_i, \mathbf{Z}_i)_{i\geqslant 1}$ be a sequence of i.i.d. copies of the $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^p$-valued random vector $(\mathbf{X}, Y, \mathbf{Z})$. Denote by f its joint density function with respect to the Lebesgue measure and by $\mathbf{g}$ the marginal density associated to the random vector $\mathbf{X}$. Parametric regression models provide powerful tools for analyzing practical data when the models are correctly specified, but may suffer from large modelling biases when structures of the models are misspecified. As an alternative, nonparametric smoothing method eases the concerns on modelling biases. However, nonparametric models are hampered by the so-called curse of dimensionality in multivariate settings, see Stone [11] for details.

One of the methods for attenuating this difficulty is to model covariate effects via a partially linear structure, a combination of linear and nonlinear parts. This results in the partially linear regression models of the form

$$Y = \mathbf{Z}^\top \beta + m(\mathbf{X}) + \varepsilon, \tag{1}$$

where $\beta \in \mathbb{R}^p$ is a vector of unknown parameters, m is the nonlinear part of the model and ε is the modelling error. Here, $\mathbf{Z}^\top$ stands as the transpose of the vector $\mathbf{Z}$.

The partially linear regression model has a broad applicability in the fields of biology, economics, education and social sciences among others. This model and various associated estimators, test statistics, and extensions have generated a substantial body of literature, which includes the works of Robinson [10], Härdle et al. [8], Liang [9], Aneiros-Pérez et al. [1] and Chen and Wang [5] together with that of Aneiros-Pérez and Vieu [2] and Dabo-Niang and Guilas [6] in the semi-functional modelling setting. Notice also the modelization of real data with such models have been carried out in a number of papers, see, for instance, Robinson [10], Härdle, Liang and Gao [8] and Aneiros-Pérez and Vieu [3].

To reduce the dimension impact of the nonparametric part in the partially linear regression model, we consider the additive structure of the regression function m and introduce the following model

$$Y = \mathbf{Z}^\top \beta + \sum_{j=1}^{d} m_j(X_j) + \varepsilon := \mathbf{Z}^\top \beta + m_{add}(\mathbf{X}) + \varepsilon, \tag{2}$$

where X_j is the j-th component of the vector $\mathbf{X}$ and m_j is a real univariate function. Subsequently, we have to estimate the unknown quantities, that is, the vector parameters β and the univariate functions $m_j, 1 \leqslant j \leqslant d$, as well.

1.1. Presentation of estimators

On the basis of the n-sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ drawn from the random variable $\mathbf{X}$, we define the kernel estimator of the marginal density $\mathbf{g}$, for any $x \in \mathbb{R}^d$, by

$$\mathbf{g}_n(\mathbf{x}) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right),$$

where K is a kernel, i.e., a non-negative function defined on $\mathbb{R}^d$ and integrating to 1, and h_n is a smoothing parameter tending to zero with a suitable rate given below. Notice that the model (1) may be written also as

$$Y - \mathbf{Z}^\top \beta = m(\mathbf{X}) + \varepsilon. \tag{3}$$

On the basis of the model (3), following the usual Nadaraya–Watson method, the regression estimator involving the nonparametric part of the model may be defined, for any $\mathbf{x} \in \mathbb{R}^d$, by

$$\widehat{m}_n^\beta(\mathbf{x}) = \sum_{i=1}^{n} \frac{Y_i - Z_i^\top \beta}{n\mathbf{g}_n(\mathbf{X}_i)} \left(\prod_{l=1}^{d} \frac{1}{h_n} K_\ell\left(\frac{x_l - X_{il}}{h_n}\right)\right), \tag{4}$$

where x_l and X_{il} are the l-th component of $\mathbf{x}$ and $\mathbf{X}_i$ respectively, and K_l $(1 \leqslant l \leqslant d)$ are kernels defined on $\mathbb{R}$. Note that $\widehat{m}_n^\beta(\mathbf{x})$ depend on the unknown parameter β which needs to be estimated. Considering the model (3), the function m clearly depends on the parameter β and its additive structure may be written as

$$m_{add}^\beta(\mathbf{x}) = \mu + \sum_{l=1}^{d} m_l^\beta(x_l), \tag{5}$$

where model identifiability considerations impose that $Em_l^\beta(X_l) = 0$, $1 \leqslant l \leqslant d$.

In order to set out our procedure, introduce first some further notations. For any $1 \leqslant l \leqslant d$, set $\mathbf{x}_{-l} = (x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_d)$, $\mathbf{q}_{-l}(\mathbf{x}_{-l}) = \prod_{j=1,\, j\neq l}^{d} q_j(x_j)$ and $\mathbf{q}(\mathbf{x}) = \prod_{l=1}^{d} q_l(x_l)$, where q_l, $1 \leqslant l \leqslant d$, are univariate densities. Following the marginal integration method, the additive regression function estimator is given, for any $\mathbf{x} \in \mathbb{R}^d$, by

$$\widehat{m}_{add}^\beta(\mathbf{x}) = \sum_{l=1}^{d} \widehat{\zeta}_l^\beta(x_l) + \int_{\mathbb{R}^d} \widehat{m}_n^\beta(\mathbf{z})\mathbf{q}(\mathbf{z})\,\mathrm{d}\mathbf{z}, \tag{6}$$

where

$$\widehat{\zeta}_l^\beta(x_l) = \int_{\mathbb{R}^{d-1}} \widehat{m}_n^\beta(\mathbf{x})\mathbf{q}_{-l}(\mathbf{x}_{-l})\,\mathrm{d}\mathbf{x}_{-l} - \int_{\mathbb{R}^d} \widehat{m}_n^\beta(\mathbf{x})\mathbf{q}(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{7}$$

Here, $\mathbf{q}_{-l}$, $1 \leqslant l \leqslant d$, and $\mathbf{q}$ stand as weight functions and $\widehat{\zeta}_l^\beta$ is the estimate of the l-th component of the additive regression function which still depends on the parameter β. Therefore, one has to estimate the vector parameter β to have ready estimates.

As a first step in the modelling procedure, we begin by the estimation of the vector parameter β. Taking the partially linear additive regression model

$$Y = \mathbf{Z}^\top \beta + m_{add}(\mathbf{X}) + \varepsilon, \tag{8}$$

making use of the statements (4), (6)–(8) and considering the least-square error criterion, it follows that

$$\widehat{\beta} = \left[\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^\top\right]^{-1}\widetilde{\mathbf{Z}}\widetilde{Y}, \tag{9}$$

where

$$\widetilde{Y} = \left[Y_i - \sum_{j=1}^{n} W_{nj}(\mathbf{X}_i)Y_j\right]_{1\leqslant i\leqslant n}, \qquad \widetilde{\mathbf{Z}} = \left[\mathbf{Z}_i - \sum_{j=1}^{n} W_{nj}(\mathbf{X}_i)\mathbf{Z}_j\right]_{1\leqslant i\leqslant n}, \qquad W_{nj}(\mathbf{X}_i) = \frac{U_{nj}(\mathbf{X}_i)}{n\mathbf{g}_n(\mathbf{X}_j)}$$

and

$$U_{nj}(\mathbf{X}_i) = \sum_{l=1}^{d} \frac{1}{h_n} K_l\left(\frac{X_{il} - X_{jl}}{h_n}\right) D_l - (d-1) \int_{\mathbb{R}^d} \prod_{k=1}^{d} \frac{1}{h_n} K_k\left(\frac{z_k - X_{jk}}{h_n}\right)\mathbf{q}(\mathbf{z})\,\mathrm{d}\mathbf{z}$$

with

$$D_l = \int\limits_{\mathbb{R}^{d-1}} \prod_{k=1,\, k\neq l} \frac{1}{h_n} K_k\left(\frac{z_k - X_{jk}}{h_n}\right) \mathbf{q}_{-l}(\mathbf{z}_{-l})\, \mathrm{d}\mathbf{z}_{-l}.$$

As a consequence, the estimates of the additive regression function and its components are defined by

$$\widehat{m}_{add}^{\widehat{\beta}}(\mathbf{z}) = \sum_{l=1}^{d} \widehat{\zeta}_l^{\widehat{\beta}}(x_l) + \int\limits_{\mathbb{R}^d} \widehat{m}_n^{\widehat{\beta}}(\mathbf{z})\mathbf{q}(\mathbf{z})\, \mathrm{d}\mathbf{z} \tag{10}$$

and

$$\widehat{\zeta}_l^{\widehat{\beta}}(x_l) = \int\limits_{\mathbb{R}^{d-1}} \widehat{m}_n^{\widehat{\beta}}(\mathbf{x})\mathbf{q}_{-l}(\mathbf{x}_{-l})\, \mathrm{d}\mathbf{x}_{-l} - \int\limits_{\mathbb{R}^d} \widehat{m}_n^{\widehat{\beta}}(\mathbf{x})\mathbf{q}(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{11}$$

2. Main results

To state our results, we consider an additional assumption on the model structure. In this respect we suppose that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^{\top} = B \quad a.s., \tag{12}$$

where B is a $p \times p$-positive definite matrix. Further assumptions involving the distribution function of $\mathbf{X}$, the regression function m, the kernels K_l, $1 \leqslant l \leqslant d$, and the smoothing parameters are gathered together hereafter for easy reference. The first part of these conditions is devoted to the regression function m and the marginal density $\mathbf{g}$. In the sequel, I^d is a compact subset of $\mathbb{R}^d$.

(G.1) m is k-times continuously differentiable.
(G.2) The marginal density $\mathbf{g}$ is strictly positive on the support I^d of the function $\mathbf{q}$.
(G.3) The marginal density $\mathbf{g}$ is uniformly continuous on its support.
(G.4) The marginal density $\mathbf{g}$ has $k+1$ continuous derivatives.

Throughout, the following hypothesis is considered upon the sequence of bandwidths $(h_n)_{n\geqslant 1}$.

(H.1) $$h_n = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{2k+1}}\right).$$

Set now, for any $\mathbf{x} \in \mathbb{R}^d$, $K(\mathbf{x}) := \prod_{l=1}^{d} K_l(x_l)$. The kernels are assumed to satisfy the following conditions:

(K.1) For any $1 \leqslant l \leqslant d$, K_l is bounded, Lipschitz continuous and integrating to one.
(K.2) For any $1 \leqslant l \leqslant d$, $K_l(u) = 0$ for $u \notin [-\lambda/2, \lambda/2]$, for some $0 < \lambda < \infty$.
(K.3) K is a kernel of order k.

Consider also the following assumptions upon the random variables Y and $\mathbf{Z}$.

(M.1) Y and $\mathbf{Z}$ are bounded.

The assumptions on the weight functions q_l, $1 \leqslant l \leqslant d$, are listed hereafter:

(Q.1) For any $1 \leqslant l \leqslant d$, q_l has $k+1$ continuous and bounded derivatives.
(Q.2) The support of the function $\mathbf{q}$ is included in the support of the density $\mathbf{g}$.

2.1. Comments on hypotheses

The most part of hypotheses are needed in considering estimation of the nonlinear part of the model to build up an estimate of the parameter β. Indeed, the proofs need to use the uniform convergence of the additive regression estimate $\widehat{m}_{add}^{\beta}$ to m_{add} given in Camlong-Viot [4] and requiring a number of conditions.

2.2. Theorems

Theorem 2.1. *Assume that assumptions* (G.1)–(G.4), (H.1), (K.1)–(K.3), (M.1), (Q.1)–(Q.2) *hold true. In addition, suppose that* $\max_{1\leqslant i\leqslant n} E|\varepsilon_i|^r < \infty$ *for some* $r \geqslant 2$. *Then, we have*

$$\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow{d} \mathcal{N}\big(0, \sigma_\varepsilon^2 B^{-1}\big), \tag{13}$$

where B^{-1} *is the inverse matrix of* B *defined above,* σ_ε^2 *is the variance of the random variable* ε *and* $\xrightarrow{d}$ *denotes the convergence in distribution.*

Theorem 2.2. *Under assumptions of Theorem* 2.1, *we have*

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\log\log n}}|\widehat{\beta}_j - \beta_j| = \big(\sigma_\varepsilon^2 b^{jj}\big)^{1/2} \quad a.s., \tag{14}$$

where $\widehat{\beta}_j$ *and* β_j *denote the* j*-th components of* $\widehat{\beta}$ *and* β *respectively and* b^{jk} *denotes the* j*-th row and* k*-th rank element of* B^{-1}.

2.3. Comments and concluding remarks

This Note is a first step to various studies and extensions related to the partially linear model when the nonlinear part is assumed to be with an additive structure and estimated using the marginal integration method. In forthcoming works, properties of estimates of the additive components of the regression function are investigated considering the exact rate of pointwise and uniform strong consistencies. Such results enable one to derive 100%-confidence bands, in the spirit of the work of Deheuvels and Mason [7], for the estimated function parameters. Notice that the asymptotic normality of these estimates will be considered and applied to build the usual $(1-\alpha)$-confidence bands for the underlying parameters. Further studies may deal with tests related to the parameters β and m and the model goodness-of-fit. Natural extensions may assume some dependency structure upon the data, as the mixing structure for example, or/and to consider functional data in the nonlinear part of the model.

2.4. Elements of proofs

Towards proving Theorem 2.1, the quantity $\sqrt{n}(\widehat{\beta}-\beta)$ is decomposed as to display a factor tending to the matrix B^{-1} times a sum of three terms where the first one is $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i\varepsilon_i$ while the two last terms are proved to be almost surely negligible. Since the first term clearly converges in distribution to a Gaussian distribution with mean zero and the covariance matrix $\sigma_\varepsilon^2 B$, the result follows. An intermediate lemma establishing the following statements

$$\max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant n}\big|W_{nj}(\mathbf{X}_i)\big| = \mathcal{O}\bigg(\frac{1}{nh}\bigg), \qquad \max_{1\leqslant i\leqslant n}\bigg|\sum_{j=1}^n W_{nj}(\mathbf{X}_i)\bigg| = \mathcal{O}(1) \quad \text{and} \quad \max_{1\leqslant i\leqslant n}\bigg|\sum_{j=1}^n W_{nj}(\mathbf{X}_i)\varepsilon_j\bigg| = \mathcal{O}\bigg(\frac{\log n}{n^{\frac{k}{2k+1}}}\bigg),$$

almost surely, has been stated. This result is helpful in proving the negligible nature of some terms.

The proof of Theorem 2.2 uses the same decomposition together with devices pertaining to the iterated logarithm law.

References

[1] G. Aneiros-Pérez, W. González-Manteiga, P. Vieu, Estimation and testing in a partial linear regression model under long-memory dependence, Bernoulli 10 (2004) 49–78.
[2] G. Aneiros-Pérez, P. Vieu, Semi-functional partial linear regression, Statist. Probab. Lett. 76 (2006) 1102–1110.
[3] G. Aneiros-Pérez, P. Vieu, Nonparametric time series prediction: a semi-functional partial linear modeling, J. Multivariate Anal. 99 (2008) 834–857.
[4] C. Camlong-Viot, Vers un test d'additivité en régression non paramétrique sous des conditions de mélange, C. R. Acad. Sci. Paris, Ser. I 333 (2001) 877–880.
[5] G. Chen, Z. Wang, The multivariate partially linear model with B-spline, Chinese J. Appl. Probab. Statist. 26 (2010) 138–150.
[6] S. Dabo-Niang, S. Guillas, Functional semiparametric partially linear model with autoregressive errors, J. Multivariate Anal. 101 (2010) 307–315.
[7] P. Deheuvels, D. Mason, General asymptotic confidence bands based on kernel-type function estimators, Stat. Inference Stoch. Process. 7 (2004) 225–277.
[8] W. Härdle, H. Liang, J. Gao, Partially Linear Models, Contributions to Statistics, Physica-Verlag, Heidelberg, 2000.
[9] H. Liang, Asymptotic normality of parametric part in partially linear models with measurement error in the nonparametric part, J. Statist. Plann. Inference 86 (2000) 51–62.
[10] P. Robinson, Root-N-consistent semiparametric regression, Econometrica 56 (1988) 931–954.
[11] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.