C. R. Acad. Sci. Paris, Ser. I 348 (2010) 207–210

Contents lists available at ScienceDirect

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

www.sciencedirect.com

Statistics

Hypothesis testing in multivariate partially linear models

L'utilisation des procédures de tests dans les modèles partiellement linéaires multidimensionnels

Marcin Przystalski

Department of Econometrics, Poznań University of Economics, Towarowa 53, 61-896 Poznań, Poland

ARTICLE INFO

Article history:
Received 28 May 2009
Accepted after revision 30 December 2009
Available online 13 February 2010

Presented by Paul Deheuvels

ABSTRACT

Multivariate partially linear models are generalizations of univariate partially linear models. In the literature, some methods of estimation of parametric and nonparametric component have been proposed. In this Note we focus on hypothesis testing of treatment effects in multivariate partially linear models. We construct a procedure for testing hypothesis H_0: $\mathbf{CBM} = \mathbf{0}$.

RÉSUMÉ

Les modèles partiellement linéaires multidimensionnels sont une généralisation des modèles partiellement linéaires unidimensionnels. Dans la littérature, on retrouve certaines méthodes d'estimation des composants paramétriques et non paramétriques. Dans cette note, nous nous concentrons sur l'utilisation des procédures de tests pour évaluer les effects du traitement dans les modèles partiellement linéaires multidimensionnels. Nous construisons une procédure pour tester l'hypothèse H_0 : $\mathbf{CBM} = \mathbf{0}$.

1. Introduction

Univariate semiparametric models have received considerable attention over the recent years (see [2,3] and the references therein) and found various practical applications e.g. in agriculture, molecular biology, econometrics and medicine. In these models the regression function can be expressed as a sum of linear and nonparametric component. In some situations, instead of using univariate models, it is necessary to model a multivariate variable. For example, in finance, it is now widely accepted that, working with series, such as asset returns, in a multidimensional framework leads to better results than working with separate univariate models. In this case, we may be interested in using a multivariate partially linear models.

Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_d)$ be $n \times d$ matrix of observations, $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n)'$ $n \times p$ design matrix, $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ is $p \times d$ matrix of unknown parameters. For each $r \in \{1, \ldots, d\}$, let f_r be an unknown function, $\boldsymbol{f}_r = (f_r(t_1), \ldots, f_r(t_n))'$, where $t_i \in D$ are known and nonrandom, $D \subset \mathbb{R}$ is a bounded domain, and $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d)$. Finally let $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d) = (\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_n)'$ be an $n \times d$ matrix of errors, where $n \geqslant p + d$. Then the multivariate partially linear model can be written as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F} + \mathbf{U}. \tag{1}$$

E-mail address: marprzyst@gmail.com.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2010.01.010

Without loss of generality, we assume that the domain $D = [0, 1]$ and for each $r \in \{1, \ldots, d\}$ f_r has $\nu \geqslant 2$ continuous derivatives on $[0, 1]$.

Pateiro-Lopéz and Gonzaléz-Manteiga [4] described estimators of $\mathbf{B}$ and $\mathbf{F}$, which generalize the estimators from a Speckman approach [6] for the multivariate case, and studied their asymptotic behaviour.

In practical situations, besides of estimation of treatment parameters, we are interested in testing hypotheses related to those parameters. In this Note, we describe a procedure for testing hypothesis H_0: $\mathbf{CBM} = \mathbf{0}$ in multivariate partially linear models.

2. Notation and assumptions

In this section we introduce some notation. We denote by bold letters $\mathbf{A}$, $\boldsymbol{a}$ matrices and vectors, respectively. $\operatorname{vec}\mathbf{A}$ will denote the vector obtained by stacking the columns of $\mathbf{A}$. We further define $\mathbf{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d)$, where $\boldsymbol{f}_r = (f_r(t_1), \ldots, f_r(t_n))$, $r = 1, 2, \ldots, d$.

Let $\mathbf{S} = (S_{n,h}(t_i, t_j))_{i,j}$, where $S_{n,h}(\cdot,\cdot)$ is a weight function depending on the bandwidth parameter h. Let $q \geqslant 1$, then for a $n \times q$ matrix $\mathbf{A}$ we write $\widetilde{\mathbf{A}} = (\mathbf{I} - \mathbf{S})\mathbf{A}$.

Let us assume, as in [4], that x_{ik}'s and t_i can be expressed as by the following regression model. Let n, $p \in \mathbb{N}$ then

$$x_{ik} = g_k(t_i) + \eta_{ik}, \tag{2}$$

where the g_k are unknown smooth functions and η_{ik} are random variables with mean zero. Using the vector notation matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \boldsymbol{g} + \boldsymbol{\eta}, \tag{3}$$

where $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{g} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p)$ with $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$, $\boldsymbol{g}_j = (g_j(t_1), \ldots, g_n(t_n))'$ and $\boldsymbol{\eta}_j = (\boldsymbol{\eta}_{1j}, \ldots, \eta_{nj})'$.

Let $i, j \in \{1, \ldots, n\}$, $k, l \in \{1, \ldots, p\}$ and $r, s \in \{1, \ldots, d\}$. Throughout the we will assume that:

(A1) The error vectors $\boldsymbol{\tau}_i$ are independent with mean vector $\mathbf{0}$ and matrix of variances and covariances $\boldsymbol{\Sigma} = (\sigma_{rs})$.
(A2) $n^{-1}\boldsymbol{\eta}'\boldsymbol{\eta} \to \mathbf{V}$, where $\mathbf{V} = (v_{ij})$ is positive definite.
(A3) $\operatorname{tr}(\mathbf{S}'\mathbf{S}) = \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^2 = \mathrm{O}(h^{-1})$.
(A4) $\|\mathbf{S}\boldsymbol{\eta}_k\|^2 = \mathrm{O}(h^{-1}) = \|\mathbf{S}'\boldsymbol{\eta}_k\|^2$.
(A5) $\tilde{g}_k(t_i) = h^{\nu} h_1(t_i) g_k^{\nu}(t_i) + \mathrm{o}(h^{\nu})$.
(A6) $\|(\mathbf{I} - \mathbf{S})\boldsymbol{f}_r\|^2 = \|\widetilde{\boldsymbol{f}}_r\|^2 = \mathrm{O}(nh^{2\nu})$.
(A7) $n^{-1}\boldsymbol{\eta}'\widetilde{\boldsymbol{f}}_r = \mathrm{O}(n^{-1/2}h^{\nu})$.
(A8) There is a probability density function $p(t)$ on $[0, 1]$ such that for each continuous function $c(t)$

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} c(t_i) = \int_0^1 c(t)p(t)\,\mathrm{d}t.$$

(A9) $\operatorname{tr}(\mathbf{S}) = \mathrm{O}(h^{-1})$.
(A10) $\max_i \sum_{j=1}^{n} |S_{ij}| = \mathrm{O}(1)$, $\max_j \sum_{i=1}^{n} |S_{ij}| = \mathrm{O}(1)$.

3. Preliminary results

Pateiro-Lopéz and Gonzaléz-Manteiga [4] proposed a method to estimate $\mathbf{B}$ and $\mathbf{F}$ in model (1), when covariance matrix of $\operatorname{vec}\mathbf{U}$ is equal to $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$. The proposed estimators $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{F}}$ generalize the Speckman approach [6] for a multivariate case and can be written as

$$\widehat{\mathbf{B}} = \left(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{Y}},$$
$$\widehat{\mathbf{F}} = \mathbf{S}(\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}).$$

In [4], Pateiro-Lopéz and Gonzaléz-Manteiga showed that the estimator $\widehat{\mathbf{B}}$ is asymptotically normal.

Theorem 3.1. *Let $h \to 0$, $nh^2 \to \infty$, and $nh^{4\nu} \to 0$ when $n \to \infty$. Suppose that either* (i) *the components of $\mathbf{X}$ are uniformly bounded or* (ii) *there is $\delta > 0$ such that, for each $i \in \{1, \ldots, n\}$ and each $k \in \{1, \ldots, p\}$ the model* (2) *holds with $E|\eta_{ik}|^{2+\delta} < C < \infty$. Then, under assumptions* (A1)–(A8) *together with* (A10)*, we have*

$$n^{1/2}\operatorname{vec}\left[\widehat{\mathbf{B}} - E(\widehat{\mathbf{B}})\right] \xrightarrow{d} N_{p\times d}\left(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{V}^{-1}\right).$$

Corollary 3.2. *Under the assumptions of Theorem* 3.1 *and the usual optimal bandwidth assumptions* ($h \sim n^{-1/(2\nu+1)}$) *we have*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_{p\times d}\big(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{V}^{-1}\big),$$

where $\widehat{\boldsymbol{\beta}} = \operatorname{vec}\widehat{\mathbf{B}}$ *and* $\boldsymbol{\beta} = \operatorname{vec}\mathbf{B}$.

Proof. Let us consider the expression $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = n^{1/2}\big[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\big] + n^{1/2}\big[E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\big].$$

By Theorem 3.1 we have that

$$n^{1/2}\big[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\big] \xrightarrow{d} N_{p\times d}\big(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{V}^{-1}\big).$$

By Theorem 2 in [4], we have that under the usual optimal bandwidth assumptions

$$n^{1/2}\big[E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}\big] \to 0.$$

This completes the proof. □

4. Main result

Suppose, one would like to test in model (1) hypothesis

$$H_0\colon \mathbf{B} = \mathbf{0}$$

or more general

$$H_0\colon \mathbf{CBM} = \mathbf{0}, \tag{4}$$

where $\mathbf{C}_{w\times p}$ is known matrix of full row rank, $w \leqslant p$, and $\mathbf{M}_{d\times q}$ is a matrix of full column rank, $q \leqslant d$. In multivariate linear regression models hypothesis (4) is tested by several test statistics, among which the most popular is Lawley–Hotelling trace statistic [5,7].

Using vec notation we can write hypothesis (4) in equivalent way $H_0' : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{L} = (\mathbf{M}' \otimes \mathbf{C})$ and $\boldsymbol{\beta} = \operatorname{vec}\mathbf{B}$. By Corollary 3.2 we have

$$n^{1/2}\mathbf{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}), \quad \text{where } \boldsymbol{\Omega} = \big(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}\big) \otimes \big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big).$$

Let us consider following statistic

$$X^2 = n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{L}'\big(\big(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}\big) \otimes \big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big)\big)^{-1}\mathbf{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Then we reject hypothesis H_0' if $X^2 > c_\alpha$, where c_α is chosen in such way that $P(X^2 > c_\alpha | H_0') = \alpha$.

Under H_0', we can simplify X^2, using properties of vec operator and Kronecker product, and we get

$$\begin{aligned} X^2 &= n\big[\operatorname{vec}(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})\big]'\big[\big(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}\big) \otimes \big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big)\big]^{-1}\big[\operatorname{vec}(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})\big] \\ &= n\operatorname{tr}\big[(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})'\big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big)^{-1}(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})\big(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}\big)^{-1}\big]. \end{aligned}$$

Because in general $\boldsymbol{\Sigma}$ is unknown, we estimate $\boldsymbol{\Sigma}$ as

$$\widehat{\boldsymbol{\Sigma}} = \frac{(n - \operatorname{tr}\mathbf{H})}{n}\mathbf{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{S} + (\mathbf{I} - \mathbf{S})\mathbf{X}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'(\mathbf{I} - \mathbf{S})$ is the hat matrix for model (2).

By the fact that $\widehat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$ and the continuous mapping theorem [8] we have

$$\widehat{\boldsymbol{\Omega}} = \big(\mathbf{M}'\widehat{\boldsymbol{\Sigma}}\mathbf{M}\big) \otimes \big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big) \xrightarrow{P} \big(\mathbf{M}'\boldsymbol{\Sigma}\mathbf{M}\big) \otimes \big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big) = \boldsymbol{\Omega}.$$

Combining this fact with the Slutsky theorem and with the central limit theorem we get

$$n^{1/2}\widehat{\boldsymbol{\Omega}}^{-1/2}\mathbf{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}).$$

Finally, we obtain that under the null hypothesis (4)

$$T_0^2 = (n - \operatorname{tr}\mathbf{H})\operatorname{tr}\big[(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})'\big(\mathbf{C}\mathbf{V}^{-1}\mathbf{C}'\big)^{-1}(\mathbf{C}\widehat{\mathbf{B}}\mathbf{M})\big(\mathbf{M}'\widehat{\boldsymbol{\Sigma}}\mathbf{M}\big)^{-1}\big] \xrightarrow{d} \chi^2_{wq}. \tag{5}$$

Remark 1. In practice matrix $\mathbf{V}$ in (5) is unknown, by Lemma 1 in [6] this matrix for sufficiently large n can be replaced by expression $n^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$.

Theorem 4.1. *Let the assumptions of Corollary* 3.2 *be satisfied. Then, under the null hypothesis* (4) *the test statistic* T_0^2 *has an asymptotic chi-square distribution with* wq *degrees of freedom.*

5. Discussion

In this Note we constructed a procedure for testing hypothesis H_0: $\mathbf{CBM} = \mathbf{0}$, based on the asymptotic result obtained by Pateiro-Lopéz and Gonzaléz-Manteiga [4]. An alternative approach to hypothesis testing problem in multivariate partially linear models can be based on profile likelihood inference proposed by Fan and Huang [1]. We would like to end this Note by stating some open questions: (i) does it exist a better approximation of the distribution of T_0^2 in model (1); (ii) can we obtain an F approximation for T_0^2 like it was proposed by McKeon (see [5,7]) for multivariate linear models?

Acknowledgements

The author wishes to thank the referee for his careful reading and valuable comments that contributed to improve the readability of the Note.

References

[1] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[2] W. Härdle, H. Liang, J. Gao, Partially Linear Models, Physica-Verlag, Würzburg, 2000.
[3] W. Härdle, M. Müller, S. Sperlich, A. Werwatz, Nonparametric and Semiparametric Models, Springer-Verlag, Berlin, Heidelberg, 2004.
[4] B. Pateiro-Lopéz, W. Gonzaléz-Manteiga, Multivariate partially linear models, Statist. Prob. Lett. 76 (2006) 1543–1549.
[5] C.R. Rao, Linear Statistical Inference and Its Applications, second ed., Wiley, New York, 1973.
[6] P. Speckman, Kernel smoothing in partial linear models, J. Roy. Stat. Soc. Ser. B 50 (1988) 413–436.
[7] N.H. Timm, Applied Multivariate Analysis, Springer-Verlag, New York, 2002.
[8] A. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.