Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 343–346
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Tests de structure en régression sur variable fonctionnelle

Laurent Delsol

Institut de mathématiques, Université de Toulouse et CNRS (UMR 5219), 118, route de Narbonne, 31062 Toulouse cedex 4, France

Reçu le 12 novembre 2007 ; accepté après révision le 29 janvier 2008

Présenté par Paul Deheuvels

Résumé

Nous proposons dans cette Note une approche générale pour la construction de tests de structure dans le cadre de la régression sur variable fonctionnelle. Comme on le fait souvent dans le cas où la variable explicative est multivariée, nous faisons apparaître dans notre statistique de test la différence entre un estimateur non-paramétrique et un estimateur particulier dépendant de l'hypothèse nulle à tester. Nous donnons la loi asymptotique de notre statistique de test sous des conditions générales, ce qui permet d'envisager l'application de notre approche dans des contextes variés. ***Pour citer cet article : L. Delsol, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abstract

Structural tests in regression on functional variables. In this Note we introduce a general approach to construct structural testing procedures in regression on functional variables. In the case of multivariate explanatory variables a well-known method consists in a comparison between a nonparametric estimator and a particular one. We adapt this approach to the case of functional explanatory variables. We give the asymptotic law of the proposed test statistic. The general approach used allows us to cover a large scope of possible applications as tests for no-effect, tests for linearity, ... ***To cite this article: L. Delsol, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

1. Introduction

Dans cet article on s'intéresse au modèle de régression usuel $Y = r(X) + \epsilon$, où Y est une variable aléatoire réelle et X une variable aléatoire à valeurs dans un espace semi-métrique $(\mathcal{E}, d)$. D'une part, de nombreux travaux ont été consacrés à l'estimation de l'opérateur de régression r dans le cas linéaire ([16,3] et [6]) et dans le cas non-paramétrique ([8,15] et [9]). D'autre part, de nombreux tests de structure ont été proposés dans le cas de variables explicatives multivariées comme en témoignent les travaux de [13,11,4] et [14]. Cependant, il semble que dans la littérature, seuls deux travaux s'intéressent à des problématiques de tests de structure en régression sur variable fonctionnelle : [2] dans le contexte particulier d'un modèle linéaire et [10] proposant un test de non-effet par des méthodes de projection (voir aussi [5] pour un test d'ajustement heuristique). Nous présentons dans cette note une méthode générale de construction de tests de structure basée sur la comparaison d'un estimateur non-paramétrique et d'un es-

Adresse e-mail : delsol@cict.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2008.01.021

timateur plus particulier, idée introduite par [13] dans le cas multivarié. Cette Note a un double objectif : présenter un résultat très général concernant la loi asymptotique de ces statistiques de test (voir paragraphe 3) et décrire plusieurs situations particulières pour lesquelles ces méthodes sont applicables (voir paragraphe 4).

2. Le modèle et la statistique de test

On considère un échantillon constitué de N paires (X_i, Y_i) indépendantes, de même loi et suivant le même modèle que (X, Y). Pour construire notre statistique de test, on décompose cet échantillon en deux échantillons indépendants notés $D = (X_i, Y_i)_{1\leqslant i\leqslant n}$ et $D^* = (X_i, Y_i)_{n+1\leqslant i\leqslant N}$ de longueurs respectives n et $m_n = N - n$.

On souhaite tester si l'opérateur de régression r a une structure particulière. Pour cela, on introduit une famille $\mathcal{R}$ d'opérateurs et on va construire une statistique de test pour tester l'hypothèse nulle

$$\mathcal{H}_0 : \big\{\exists r_0 \in \mathcal{R},\, P\big(r(X) = r_0(X)\big) = 1\big\}$$

contre l'alternative locale

$$\mathcal{H}_{1,n} : \Big\{ \inf_{r_0\in\mathcal{R}} \|r - r_0\|_{\mathbb{L}^2(w\,\mathrm{d}P_X)} \geqslant \eta_n \Big\},$$

où w est une fonction de poids et $\mathrm{d}P_X$ est la loi de X. La manière dont la suite η_n décroit vers 0 reflète la capacité de détecter des différences de plus en plus fines entre r et la famille $\mathcal{R}$ lorsque n croît. On s'inspire de [13] et nos résultats pourraient être adaptés à des statistiques de test de même nature. Cependant, des raisons de facilité de compréhension et de généralité nous incitent à considérer une statistique de test différente (introduite par [11] dans le cas multivarié) qui s'écrit :

$$T_n^* = \int_{\mathcal{E}} \left(\sum_{i=1}^{n} \big(Y_i - r_0^*(X_i)\big) K\left(\frac{d(x, X_i)}{h_n}\right)\right)^2 w(x)\,\mathrm{d}P_X(x),$$

dans laquelle K est un noyau à support compact, h_n est un paramètre de lissage et r_0^* est un estimateur particulier, adapté à l'hypothèse nulle $\mathcal{H}_0$, construit à partir de l'échantillon D^*. Estimer r_0^* à partir de D^* simplifie l'étude asymptotique et rend plus lisibles les conditions portant sur r_0^*. On peut cependant envisager d'étudier le cas où le même échantillon sert pour construire r_0^* et T_n^*.

3. Résultat général

3.1. Notations et hypothèses

On commence par faire des hypothèses concernant la statistique de test :

$$w \text{ est positive et bornée sur son support borné } W \text{ et } \mathbb{P}(w(X) = 0) \neq 1. \tag{1}$$

$$K \text{ a pour support le compact } [0,\,1],\ \text{est décroissante et } \mathcal{C}^1 \text{ sur }]0,\,1[\text{ et } K(1) > 0. \tag{2}$$

On continue, en notant $W_\gamma = \{x \in \mathcal{E},\, \exists y \in W,\, d(x, y) < \gamma\}$, avec des hypothèses portant sur le modèle :

$$\exists \alpha > 0,\ r \text{ est Hölderien d'ordre } \beta \text{ sur } W_\alpha,\ \text{par rapport à } d. \tag{3}$$

$$\exists M > 0, \mathbb{E}[\epsilon^4|X] \leqslant M \text{ p.s. et } \mathbb{E}[\epsilon^2|X] \equiv \sigma_\epsilon^2 > 0. \tag{4}$$

On a également besoin d'introduire les notations suivantes :

$$F_x(s) = \mathbb{P}\big(d(x, X) \leqslant s\big), \qquad F_{x,y}(s, t) = \mathbb{P}\big(d(x, X) \leqslant s, d(y, X) \leqslant t\big),$$

$$\Omega_1(s) = \int_{\mathcal{E}} F_x(s) w(x)\,\mathrm{d}P_X(x), \qquad \Omega_2(s) = \int_{\mathcal{E}\times\mathcal{E}} F_{x,y}^2(s, s) w(x) w(y)\,\mathrm{d}P_X(x)\,\mathrm{d}P_X(y).$$

Viennent à présent des hypothèses reliant la suite η_n, le paramètre de lissage h_n et les probabilités de petites boules :

$$\exists C_1, C_2, \gamma > 0, \exists \Phi, \forall x \in W_\gamma, \forall n \in \mathbb{N}, \quad C_1 \Phi(h_n) \leqslant F_x(h_n) \leqslant C_2 \Phi(h_n), \tag{5}$$

$$\exists C > 0, \theta_n := C v_n \left(\frac{1}{n^{\frac{1}{2}} \Phi^{\frac{1}{4}}(h_n)} + h_n^\beta \right) \leqslant \eta_n, \quad \text{avec } v_n \to +\infty \text{ et } \theta_n \to 0, \tag{6}$$

$$\exists C_3 > 0, \Omega_2(h_n) \geqslant C_3 \Phi^{3+l}(h_n) \quad \text{avec } l < \frac{1}{2} \text{ et } n\Phi^{1+2l}(h_n) \to +\infty. \tag{7}$$

Pour toute famille $\mathcal{G}$ convexe non-vide de $\mathbb{L}^2(\mathrm{d}P_X)$ on notera $\overline{\mathcal{G}}$ son adhérence et pour tout opérateur g de $\mathbb{L}^2(\mathrm{d}P_X)$ on notera la distance par projection de g à $\overline{\mathcal{G}}$ par $d_{\overline{\mathcal{G}}}(g)$. On suppose que la famille $\mathcal{R}$ possède les propriétés suivantes :

$$\mathcal{R} \text{ est non-vide et convexe}, \exists \alpha > 0, \mathcal{R} \subset \mathbb{L}^2_{(\mathrm{d}P_X)}(W_\alpha), \tag{8}$$

afin de pouvoir définir la projection r_0 de r sur $\overline{\mathcal{R}}$. On note $\mathbb{E}_*$ l'espérance conditionnelle à D^* et $Hol(C, \beta)$ l'ensemble des fonctions Hölderiennes d'ordre β et de constante C. On demande enfin que l'estimateur r_0^* vérifie les hypothèses suivantes :

$$\text{sous } \mathcal{H}_0, \quad n\Phi^{\frac{1-l}{2}}(h_n)\mathbb{E}_*\left[\left(r_0^*(X_1) - r_0(X_1)\right)^2 1_{W_\alpha}(X_1)\right] \xrightarrow{P} 0, \tag{9}$$

$$\text{sous } \mathcal{H}_{1,n}, \quad \eta_n^{-2} d_F^2(r_0^*) \xrightarrow{P} 0 \text{ et } \exists C > 0, \eta_n^{-2} d^2_{Hol(C,\beta)}(r_0^*) \xrightarrow{P} 0, \tag{10}$$

$$\text{sous } \mathcal{H}_{1,n}, \quad \mathbb{E}_*\left[\left(r_0^*(X)\right)^4 1_{W_\alpha}(X)\right] = O_p(1). \tag{11}$$

On pourra choisir l'ordre de grandeur de m_n par rapport à n afin que ces hypothèses soient vérifiées.

3.2. Loi asymptotique de T_n^*

Pour exprimer la loi asymptotique de T_n^* sous $\mathcal{H}_0$, on introduit $T_{1,n}$ et $T_{2,n}$ qui permettent d'écrire les termes de biais et de variance.

$$T_{1,n} = \int_{\mathcal{E}} \sum_{i=1}^n K^2\left(\frac{d(X_i, x)}{h_n}\right) \epsilon_i^2 w(x)\, \mathrm{d}P_X(x),$$

$$T_{2,n} = \int_{\mathcal{E}} \sum_{1 \leqslant i \neq j \leqslant n} K\left(\frac{d(X_i, x)}{h_n}\right) K\left(\frac{d(X_j, x)}{h_n}\right) \epsilon_i \epsilon_j w(x)\, \mathrm{d}P_X(x).$$

A partir des hypothèses précédentes il est possible de montrer la normalité asymptotique de notre statistique de test sous $\mathcal{H}_0$ ainsi que sa divergence sous $\mathcal{H}_{1,n}$. C'est l'objet du théorème suivant, dont la preuve est donnée, sous différents jeux d'hypothèses, dans [7].

Théorème 3.1. *Sous les hypothèses* (1)–(11) *on a* :

- *Sous* $(\mathcal{H}_0)$, $\frac{1}{\sqrt{Var(T_{2,n})}}(T_n^* - \mathbb{E}[T_{1,n}]) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$;
- *Sous* $(\mathcal{H}_{1,n})$, $\frac{1}{\sqrt{Var(T_{2,n})}}(T_n^* - \mathbb{E}[T_{1,n}]) \xrightarrow{\mathcal{L}} +\infty$.

Dans le cas particulier où $K \equiv 1_{[0;1]}$, les expressions asymptotiques de l'espérance et de la variance sont les suivantes :

$$\mathbb{E}[T_{1,n}] = n\Omega_1(h_n)\sigma_\epsilon^2 \quad \text{et} \quad Var(T_{2,n}) \sim 2n^2(\sigma_\epsilon^2)^2\Omega_2(h_n).$$

Pour d'autres valeurs de K leurs expressions sont plus complexes et sont détaillées dans [7].

Idée de la preuve. On part d'une décomposition de la statistique T_n^*. On obtient la normalité asymptotique de $T_{2,n}$ grâce à un CLT pour U-statistiques donné par [12]. Ensuite, on compare les ordres de grandeur des moments des autres termes. □

4. Quelques cas particuliers importants

La complexité des hypothèses introduites pour établir le Théorème 3.1 est essentiellement liée à la volonté de considérer des situations aussi générales que possible, afin d'accroître au maximum les possibilités d'application. Afin d'insister sur la faible restrictivité de ces hypothèses, et notamment celles portant sur r_0^*, voici quelques exemples de situations variées dans lesquelles le Théorème 3.1 ou une de ses généralisations données dans [7] est applicable.

- Si $\mathcal{R} = \{r_0\}$, on peut appliquer le Théorème 3.1 avec $r_0^* = r_0$ et $n = N$ lorsque r_0 est Hölderien sur un voisinage de W. Ceci complète les travaux de [2] (modèle linéaire fonctionnel) at de [10] (méthodes de projection).
- Si $\mathcal{R} = \{r_0, \exists C, r_0 \equiv C\}$, on peut utiliser notre statistique de test avec $r_0^* \equiv m_n^{-1} \sum_{i=n+1}^{N} Y_i$ si $n\Phi^{\frac{1-l}{2}}(h_n) = o(m_n)$. On obtient ainsi un test de non-effet de X sur Y. Ceci complète les approches introduite par [2] et [10].
- Si $\mathcal{R} = \{r_0 : \mathcal{E} \to \mathbb{R},\ \text{linéaire}\}$, on peut appliquer, sous de bonnes hypothèses, nos résultats en prenant pour r_0^* l'estimateur étudié dans [6]. On obtient ainsi le premier test de linéarité proposé pour un modèle de régression sur variable fonctionnelle.
- Soit $V : \mathcal{E} \to \mathbb{R}^q$ connu. Si $\mathcal{R} = \{r_0, \exists \psi : \mathbb{R}^q \to \mathbb{R}, r_0 = \psi \circ V\}$, on peut appliquer, sous de bonnes hypothèses, nos résultats en prenant pour r_0^* l'estimateur à noyau construit à partir de $(V(X_i), Y_i)_{n+1 \leqslant i \leqslant N}$. On obtient alors un test innovant qui permet de tester si l'effet de la variable explicative fonctionnelle est réductible à l'effet d'un vecteur $V(X)$ constitué de valeurs caractéristiques de cette courbe (minima, maxima, points d'inflexion, ...).
- Si $\mathcal{E}$ est un espace de Hilbert et si $\mathcal{R} = \{r_0, \exists \theta \in \mathcal{E}, \exists \psi : \mathbb{R} \to \mathbb{R}, r_0 = \psi(\langle \cdot, \theta \rangle)\}$, les résultats donnés par [1] indiquent que l'on peut déterminer θ par validation croisée (θ_{CV}) et prendre pour r_0^* l'estimateur à noyau appliqué aux variables $(\langle X_i, \theta_{CV} \rangle, Y_i)_{n+1 \leqslant i \leqslant N}$.

Des travaux sont actuellement en cours afin, notamment, de régler des détails de mise en oeuvre sur données réelles. On peut penser généraliser notre approche à d'autres tests de structure permettant par exemple de tester si le modèle est semi-paramétrique, paramétrique, additif, ... Suivant le test de structure que l'on souhaite réaliser, le problème clé consiste à trouver un estimateur r_0^* qui vérifie les hypothèses (9)–(11) ou des variantes proposées dans [7].

Remerciements

Je remercie le rapporteur pour ses questions et commentaires pertinents qui ont permis d'améliorer cette Note. Je tiens à exprimer ma reconnaissance au groupe STAPH et plus particulièrement à Frédéric Ferraty et Philippe Vieu.

Références

[1] A. Ait-Saïdi, F. Ferraty, R. Kassa, P. Vieu, Cross-validated estimations in the single functional index model, Statistics (2008), in press.
[2] H. Cardot, F. Ferraty, A. Mas, P. Sarda, Testing hypotheses in the functional linear model, Scand. J. Statist. 30 (2003) 241–255.
[3] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statist. Sinica 13 (3) (2003) 571–591.
[4] S.X. Chen, I. Van Keilegom, A goodness-of-fit test for parametric and semiparametric models in multiresponse regression, Inst. de Statistique, U.C.L., Discussion paper 0616, 2006.
[5] J.M. Chiou, H.-G. Müller, Diagnostics for functional regression via residual processes, Comput. Statist. Data Anal. 51 (10) (2007) 4849–4863.
[6] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, Ann. Statist. (2008), in press.
[7] L. Delsol, F. Ferraty, P. Vieu, Structural testing procedures in regression on functional variables, (2008), submitted for publication.
[8] F. Ferraty, A. Goia, P. Vieu, Functional nonparametric model for time series: a fractal approach for dimension reduction, Test 11 (2) (2002) 317–344.
[9] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer-Verlag, New York, 2006.
[10] D. Gadiaga, R. Ignaccolo, Test of no-effect hypothesis, Afrika Stat. 1 (1) (2005) 67–76.
[11] W. González-Manteiga, A. Quintela-del-Río, P. Vieu, A note on variable selection in nonparametric regression with dependent data, Statist. Probab. Lett. 57 (2002) 259–268.
[12] P. Hall, Central limit theorem for integrated square error of multivariate nonparametric density estimators, J. Multivariate Anal. 14 (1) (1984) 1–16.
[13] W. Härdle, E. Mammen, Comparing nonparametric versus parametric regression fits, Ann. Statist. 21 (4) (1993) 1926–1947.
[14] P. Lavergne, V. Patilea, Un test sur la significativité des variables explicatives en régression non-paramétrique, JDS Angers 2007, 2007.
[15] E. Masry, Nonparametric regression estimation for dependent functional data: asymptotic normality, Stochastic Process. Appl. 115 (1) (2005) 155–177.
[16] J. Ramsay, C. Dalzell, Some tools for functional data analysis, J. R. Statist. Soc. B. 53 (1991) 539–572.