Available online at www.sciencedirect.com
SCIENCE @ DIRECT®
ELSEVIER
C. R. Acad. Sci. Paris, Ser. I 336 (2003) 601–604
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Une méthode semi-paramétrique pour tester un modèle de régression
A semi-parametric method to test a regression model

Michel Harel [a,b]

[a] *IUFM du Limousin (et UMRC 55830, CNRS, Toulouse), 209, bd de Vanteaux, 87036 Limoges cedex, France*
[b] *Centro de Modelamentio UMR CNRS 2071, Universidad de Chile, Santiago, Chile*

Reçu le 24 octobre 2002 ; accepté après révision le 23 janvier 2003

Présenté par Paul Deheuvels

Résumé

Le but est de tester l'hypothèse H_0 qu'un modèle de régression est paramétrique et appartient à une famille donnée contre l'alternative H_1 approchant l'hypothèse dans une direction spécifique au taux $n^{-1/2}$. Pour cela, nous considérons un processus empirique tel que sous l'hypothèses H_0 ce processus dépend d'un paramètre θ_0. D'abord, nous commençons par estimer le paramètre et nous montrons que le processus empirique converge en loi vers un certain processus Gaussien si le paramètre est remplacé par son estimateur $\widetilde{\theta_n}$. Cependant il est important de vérifier l'impact d'une alternative qui approche H_0 dans une direction spécifique (au taux $n^{1/2}$). Pour cela, nous avons besoin de tests qui soient consistants sur toute l'alternative H_1. Notre idée est d'utiliser un processus empirique marqué basé sur les résidus qui converge en loi vers un processus Gaussien. ***Pour citer cet article : M. Harel, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

Abstract

The purpose is to test the hypothesis H_0 that a regression model is parametric and belongs to a given family versus the alternative H_1 approaches the hypothesis from a specific direction at the rate $n^{-1/2}$. For that, we consider an empirical process such that under H_0 this process depends of a parameter θ_0. First, we start by estimating the parameter and we prove that the empirical process converges in distribution to a certain Gaussian process when the parameter is replaced by its estimator $\widetilde{\theta_n}$. However it is important to check the impact of an alternative approaching H_0 from a specific direction (at the rate $n^{1/2}$). For that, we need tests which are consistent on the whole of H_1. Our idea is to use a marked empirical process based on residuals which converges in distribution to a Gaussian process. ***To cite this article: M. Harel, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

1. Introduction

Soit $\{\mathbf{Z}_i = (\mathbf{X}_i, Y_i);\ i \geqslant 1\}$ une suite de variables aléatoires de fonctions de répartition continues $H(\mathbf{z})$, $\mathbf{z} \in \mathbb{R}^{1+d}$, et nous supposons que $H(\mathbf{z})$ admet une densité strictement positive et H possède deux lois marginales F et G (respectivement de $\mathbf{X}_i$ et Y_i) où les densités respectives sont notées f et g.

Adresse e-mail : harel@unilim.fr (M. Harel).

1631-073X/03/$ – see front matter © 2003 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS. Tous droits réservés.
doi:10.1016/S1631-073X(03)00128-6

Supposons que les variables aléatoires Y_i sont intégrables de sorte que la fonction de régression $m(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$ de Y sur $\mathbf{X}$ is est bien définie où $\mathbf{x} \in \mathbb{R}^d$, $(\mathbf{X}, Y)$ a la même distribution que $(\mathbf{X}_i, Y_i)$ et est p.s. en $\mathbf{x}$ uniquement définie du point de vue de l'équation

$$m(\mathbf{X}) = E(Y \mid \mathbf{X}). \tag{1}$$

Dans cette Note, nous supposons que la suite $\{\mathbf{Z}_i\}$ est absolument régulière avec un taux géométrique (pour une définition voir Harel et Puri [3]).

Dans beaucoup de littérature, on considère des modèles paramétriques de telle sorte que m est supposé appartenir à une famille donnée

$$\mathcal{H} = \bigl\{m(\cdot, \theta),\ \theta \in \Theta\bigr\} \tag{2}$$

de fonctions, où $\theta \subset \mathbb{R}^p$ est un ensemble de paramètres.

Nous supposons que $m(\mathbf{x}) = m(\mathbf{x}, \theta_0)$ pour la vraie valeur du paramètre θ_0. Le problème est comment estimer θ_0 ou tester l'hypothèse que la valeur du paramètre est θ_0. Un cas bien connu est le modèle linéaire pour lequel $m(\mathbf{x}, \theta) = g'(\mathbf{x})\theta$, g est une fonction vectorielle connue. Beaucoup de travaux ont été réalisés sur la manière d'estimer m d'une façon complètement nonparamétrique voir Stute [5,6]) si la suite $\{\mathbf{Z}_i\}$ est i.i.d. ensuite Yoshihara [8] et Harel et Puri [3] si la suite $\{\mathbf{Z}_i\}$ est absolument régulière.

Il est important de vérifier l'impact d'une alternative qui approche H_0 dans une direction spécifique (au taux $n^{1/2}$). Pour cela, nous avons besoin de tests qui soient consistants sur tout H_1. Notre idée est d'utiliser un processus empirique marqué basé sur les résidus convergeant en loi vers un processus Gaussien. Sous l'hypothèse H_0, nous supposons que la vraie valeur m est égale à $m(\cdot, \theta_0)$ et soit $\widetilde{\theta_n}$ un estimateur de θ_0, alors nous rejetons l'hypothèse si le processus empirique marqué (sur lequel θ_0 est remplacé par $\widetilde{\theta_n}$) excède une valeur critique.

Maintenant définissons

$$R_n^*(\mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} I_{[\mathbf{X}_i \leqslant \mathbf{x}]}\bigl(Y_i - m(\mathbf{X}_i, \widetilde{\theta_n})\bigr), \quad \mathbf{x} \in \mathbb{R}^d, \tag{3}$$

le processus empirique marqué.

Le résultat principal montrera la convergence faible du processus R_n^* par rapport à la topologie de Skorohod sous des conditions raisonnables. De tels résultats ont été donnés par Stute [7] seulement quand la suite $\{\mathbf{Z}_i\}$ est i.i.d., des applications sont données pour des modèles linéaires, son estimateur $\widetilde{\theta_n}$ de θ_0 est l'estimateur des moindres carrés et son espace alternatif est un sous espace de L^2 (l'espace des fonctions de carré intégrable) puis par Diebold et Zuber [1]. Notre hypothèse d'absolue régularité géométrique nous donnera des applications pour des modèles plus généraux comme les modèles autorégressifs. Notre estimateur $\widetilde{\theta_n}$ de θ_0 sera le α-quantile d'autorégression qui est beaucoup plus robuste si les bruits blancs ne sont pas nécessairement des lois normales. Le cas des modèles autorégressifs a été abordé par Koul et Stute [4] avec un estimateur classique de θ_0 et la consistance de l'alternative n'est pas abordée.

2. Conditions et convergence faible du processus empirique marqué

Sans perte de généralité, nous supposons maintenant que $d = 1$.

Nous supposons que la suite $\{\mathbf{Z}_i = (X_i, Y_i),\ i \geqslant 1\}$ est absolument régulière avec le taux

$$\beta(m) = 0\bigl(\rho^m\bigr), \quad 0 < \rho < 1. \tag{4}$$

Nous savons que le processus R_n^* défini dans (3) prend ses valeurs dans l'espace de Skorohod $D(-\infty, +\infty)$. Pour traiter de telles statistiques, nous étendons de façon continue R_n^* à $-\infty$ et $+\infty$ en posant $R_n^*(-\infty) = 0$ et $R_n^*(+\infty) = n^{-1/2} \sum_{i=1}^n (Y_i - m(\mathbf{X},\ \widetilde{\theta_n}))$. Alors R_n^* devient un processus à valeurs dans $D[-\infty, +\infty]$.

Pour le comportement du processus R_n^*, certaines conditions de régularité sur l'estimateur $\widetilde{\theta_n}$ seront nécessaires. Ces conditions sont similaires aux conditions de Stute [7] et nous ne les appliquerons pas à l'estimateur des moindres carrés mais à notre α-quantile d'autorégression et notre suite $\{\mathbf{Z}_i\}$ n'est pas i.i.d. mais absolument régulière.

Condition 1. Sous H_0, c'est à dire $m = m(\cdot, \boldsymbol{\theta}_0)$ pour un certain $\boldsymbol{\theta}_0 \in \Theta$, inconnu, $\widetilde{\boldsymbol{\theta}_n}$ admet un développement

$$n^{1/2}\left(\widetilde{\boldsymbol{\theta}_n} - \boldsymbol{\theta}_0\right) = n^{-1/2} \sum_{i=1}^{n} \mathbf{l}(X_i, Y_i, \boldsymbol{\theta}_0) + \mathrm{o}_p(1)$$

pour une certaine fonction vectorielle $\mathbf{l}$ telle que (i) $E[\mathbf{l}(X, Y, \boldsymbol{\theta}_0)] = 0$; (ii) $L_{i,j}(\boldsymbol{\theta}_0) = E[\mathbf{l}(X_i, Y_i, \boldsymbol{\theta}_0)\mathbf{l}'(X_j, Y_j, \boldsymbol{\theta}_0)]$ existe pour tout $(i, j) \in (\mathbb{N}^*)^2$.

Condition 2. (i) $m(x, \boldsymbol{\theta})$ est continuement différentiable pour chaque $\boldsymbol{\theta}$ dans l'intérieur Θ^0 de Θ. Posons

$$\mathbf{g}(x, \theta) = \frac{\partial m(x, \boldsymbol{\theta})}{\partial \theta} = \left(g_1(x, \boldsymbol{\theta}), \ldots, g_p(x, \boldsymbol{\theta})\right)'. \tag{5}$$

(ii) Il existe une fonction F-intégrable $M(x)$ telle que

$$\left|g_i(x, \boldsymbol{\theta})\right| \leqslant M(x) \quad \text{pour tout } \boldsymbol{\theta} \in \Theta \text{ et } 1 \leqslant i \leqslant p. \tag{6}$$

Théorème 1. *Supposons que $E(Y^{2+\delta_0}) < +\infty$ où $\delta_0 > 0$ et que les Conditions 1 and 2 soient satisfaites. Alors sous $m = m(\cdot, \boldsymbol{\theta}_0)$, $R_n^* \to R_\infty^*$ en loi dans l'espace $D[-\infty, +\infty]$ où R_∞^* est un processus centré Gaussien avec fonction de covariance*

$$\begin{aligned} K^*(x, y) \ = \ & K(x, y) + \mathbf{G}'(x, \boldsymbol{\theta}_0)\left(L_{1,1}(\boldsymbol{\theta}_0) + 2\sum_{k=2}^{+\infty} L_{1,k}(\theta_0)\right) G(y, \boldsymbol{\theta}_0) \\ & - \mathbf{G}'(x, \boldsymbol{\theta}_0) \sum_{k=0}^{+\infty} E\left[I_{[X_1 \leqslant x]}\left(Y_1 - m(X_1, \boldsymbol{\theta}_0)\right)\mathbf{l}(X_{1+k}, Y_{1+k}, \boldsymbol{\theta}_0)\right] \\ & - \mathbf{G}'(y, \boldsymbol{\theta}_0) \sum_{k=0}^{+\infty} E\left[I_{[X_1 \leqslant y]}\left(Y_1 - m(X_1, \boldsymbol{\theta}_0)\right)\mathbf{l}(X_{1+k}, Y_{1+k}, \boldsymbol{\theta}_0)\right], \end{aligned} \tag{7}$$

où $\mathbf{G}(x, \boldsymbol{\theta}) = (G_1(x, \boldsymbol{\theta}), \ldots, G_p(x, \boldsymbol{\theta}))$ et $G_i(x, \boldsymbol{\theta}) = \int_{-\infty}^{x} g_i(u, \theta) F(\mathrm{d}u)$, $1 \leqslant i \leqslant p$, et

$$K(x, y) \ = \int_{-\infty}^{x \wedge y} \mathrm{Var}(Y \mid X = u) F(\mathrm{d}u) + 2\sum_{k=1}^{+\infty} \int_{-\infty}^{x} \int_{-\infty}^{y} \mathrm{Cov}(Y_1, Y_{1+k} \mid X_1 = u, X_{1+k} = v) F_{1,1+k}(\mathrm{d}u, \mathrm{d}v), \tag{8}$$

où $F_{i,j}$ est la fonction de répartition de (X_i, X_j).

3. Applications au modèle d'autorégression

Considérons un modèle qui peut être écrit sous la forme

$$X_i^* = \rho_0 + \rho_1 X_{i-1}^* + \cdots + \rho_p X_{i-d}^* + \varepsilon_i, \tag{9}$$

où $\boldsymbol{\rho} = (\rho_0, \ldots, \rho_d)$ est un paramètre d'intérêt avec les (X_i^*, ε_i) identiquement distribué et géométriquement absolument réguliers.

Les ε_i peuvent dépendre de X_{i-1}^* contrairement aux processus ARMA.

Nous supposons aussi que

(i) $E(\|\varepsilon_i\|^5) < +\infty$,
(ii) toute les racines de l'équation $x^p - \rho_1 x^{p-1} - \cdots - \rho_d = 0$ sont à l'intérieur du cercle unité,
(iii) la densité f^* de F^* (fonction de répartition de ε_i) est bornée, strictement positive et différentiable avec $f^{*'}$ absolument bornée.

Soit $(\mathbf{X}_i, Y_i)$ la suite de vecteurs aléatoires dans $\mathbb{R}^{d+1}$ définis par $Y_i = X_i^*$ et $\mathbf{X}_i = (X_{i-1}^*, \ldots, X_{i-d}^*)$.

Posons maintenant $\mathcal{M} = \{m, m(\mathbf{x}) = E(Y_i \mid \mathbf{X}_i = \mathbf{x})$ et X_i^* est de la forme (9)$\}$.

Nous utiliserons les résultats de la Section 2 pour tester $H_0 : m \in \mathcal{M}$ versus la suite d'alternatives $H_{1,n} : m \equiv m_n \notin \mathcal{M}$.

Nous pouvons obtenir la puissance des tests si l'alternative approche $\mathcal{M}$ au taux $n^{-1/2}$ dans une direction spécifique.

Le paramètre $\boldsymbol{\rho}$ sera estimé par le $\alpha^{\text{ème}}$ quantile d'autorégression (voir Harel et Puri [2] pour une définition).

D'après le Théorème 3.1 de Harel et Puri [2], le $\alpha^{\text{ème}}$ quantile d'autorégression satisfait la Condition 1 et le modèle (9) satisfait la Condition 2.

Notons $\mathbf{D}_d[-\infty, +\infty]$ la version d-dimensionnelle de l'espace de Skorohod $D[-\infty, +\infty]$.

Nous déduisons facilement le théorème suivant

Théorème 2. *Sous H_0 et les conditions* (i)–(iii), *$R_n^* \to R_\infty^*$ en loi dans l'espace $\mathbf{D}_d[-\infty, +\infty]$ où R_∞^* est un processus centré Gaussien avec pour fonction de covariance*

$$\begin{aligned} K^*(\mathbf{x},\mathbf{y}) &= K(\mathbf{x},\mathbf{y}) + \mathbf{G}'(\mathbf{x},\boldsymbol{\rho})\mathbf{Q}_\alpha^{-1}\left\{E\left(V_1^2(\alpha)\mathbf{X}_1^{*'}\mathbf{X}_1^*\right) + 2\sum_{k=1}^{+\infty} E\left(V_1(\alpha)V_{1+k}(\alpha)\mathbf{X}_1^{*'}\mathbf{X}_{1+k}^*\right)\right\}\mathbf{Q}_\alpha^{-1}\mathbf{G}(\mathbf{y},\boldsymbol{\rho}) \\ &\quad - \mathbf{G}'(\mathbf{x},\boldsymbol{\rho})\mathbf{Q}_\alpha^{-1}\sum_{k=0}^{+\infty} E\left[I_{[\mathbf{X}_1\leqslant\mathbf{x}]}\varepsilon_1 V_{1+k}(\alpha)\mathbf{X}_1^{*'}\mathbf{X}_{1+k}^*\right] - \mathbf{G}'(\mathbf{y},\boldsymbol{\rho})\mathbf{Q}_\alpha^{-1}\sum_{k=0}^{+\infty} E\left[I_{[\mathbf{X}_1\leqslant\mathbf{y}]}\varepsilon_1 V_{1+k}(\alpha)\mathbf{X}_1^{*'}\mathbf{X}_{1+k}^*\right], \end{aligned} \tag{10}$$

où $V_i(\alpha) = I_{[F^*(\varepsilon_i)\leqslant\alpha]} - \alpha$, $\mathbf{G}'(\mathbf{x},\boldsymbol{\rho}) = (G_0(\mathbf{x},\boldsymbol{\rho}), G_1(\mathbf{x},\boldsymbol{\rho}), \ldots, G_d(\mathbf{x},\boldsymbol{\rho}))$, $G_j(\mathbf{x},\rho) = E[I_{[\mathbf{X}_1\leqslant\mathbf{x}]}\mathbf{X}_1^{*'}\mathbf{e}_j]$ et $\mathbf{Q}_\alpha$ est une $(d+1)\times(d+1)$ matrice définie positive.

Maintenant nous avons besoin de critères pour tester H_0 contre $\mathrm{H}_{1,n}$.

Supposons que sous $\mathrm{H}_{1,n}$ le modèle est

$$\mathbf{Y}_n = \boldsymbol{\xi}_n\boldsymbol{\rho} + \mathbf{V}_n\boldsymbol{\beta}_n + \boldsymbol{\varepsilon}_n, \tag{11}$$

où $\mathbf{V}_n = (v_{j,i})_{\substack{1\leqslant j\leqslant n\\ 1\leqslant i\leqslant q}}$ est une matrice d'ordre $(n\times q)$, $\boldsymbol{\beta}_n \in \mathbb{R}^q$ non spécifié, $\mathbf{Y}_n' = (X_1, \ldots, X_n)$ et $\boldsymbol{\xi}_n$ est la $n\times(d+1)$ matrice dont la $\mathrm{i}^{\text{ème}}$*colonne* est $(1, X_{i-1}, \ldots, X_{i-d})'$.

Nous supposons aussi que (iv) $\lim_{n\to\infty} n^{-1}\sum_{j=1}^{i=n} v_{j,i} = v_i$, où $-\infty < v_i < +\infty$, $1\leqslant i\leqslant q$ (posons $\mathbf{V} = (v_1, \ldots, v_q)'$), (v) $\boldsymbol{\beta}_n = (\beta_{n,1}, \ldots, \beta_{n,q}) = n^{-1/2}\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_0 \in \mathbb{R}^q$, fixé.

On déduit facilement du Théorème 2

Théorème 3. *Sous $H_{1,n}$ les conditions* (i) *à* (v), *$R_n^* \to R_\infty^*$ en loi dans l'espace $\mathbf{D}_d[-\infty, +\infty]$ où R_∞^* est un processus Gaussien de moyenne $\mathbf{V}'\boldsymbol{\beta}_0$ et de fonction de covariance $K^*(\mathbf{x},\mathbf{y})$ définie dans* (10).

Références

[1] J. Diebold, J. Zuber, Goodness of fit tests for nonlinear heteroscedic regression models, Statist. Probab. Lett. 42 (1999) 53–60.

[2] M. Harel, M.L. Puri, Autoregression quantiles and related rank score processes for generalized random coefficient autoregressive processes, J. Statist. Plann. Inference 68 (1998) 271–294.

[3] M. Harel, M.L. Puri, Conditional empirical processes defined by nonstationary absolutely regular sequences, J. Multivariate Anal. 70 (1999) 250–285.

[4] H.L. Koul, W. Stute, Nonparametric model checks for time series, Ann. Statist. 27 (1999) 204–236.

[5] W. Stute, Asymptotic normality on nearest neighbor regression functions estimates, Ann. Statist. 12 (1984) 917–926.

[6] W. Stute, On almost sure convergence of conditional empirical distribution functions, Ann. Probab. 14 (1986) 891–901.

[7] W. Stute, Nonparametric model checks for regression, Ann. Statist. 25 (1997) 613–641.

[8] K. Yoshihara, Conditional empirical processes defined by φ-mixing sequences, Comput. Math. Appl. 19 (1990) 149–158.