Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 341 (2005) 313–316
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Un test d'adéquation global pour la fonction de répartition conditionnelle

Sandie Ferrigno, Gilles R. Ducharme

Équipe probabilités et statistique, institut de mathématiques et de modélisation de Montpellier, UMR CNRS 5149, CC051, université Montpellier-II, place E.Bataillon, 34095 Montpellier cedex 05, France

Reçu le 25 mars 2005 ; accepté après révision le 29 juin 2005

Disponible sur Internet le 24 août 2005

Présenté par Paul Deheuvels

Résumé

Nous proposons un test d'adéquation global permettant de juger de la validité d'un modèle statistique en testant simultanément toutes les suppositions sous-jacentes à celui-ci. Ce test est basé sur l'estimateur polynomial local de la fonction de répartition conditionnelle et sur un paradigme standard qui mesure la distance entre cet estimateur non paramétrique et le modèle paramétrique théorique. Nous obtenons le comportement asymptotique de la statistique du test à la fois sous l'hypothèse nulle et sous des alternatives locales. ***Pour citer cet article : S. Ferrigno, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 341 (2005).***

Abstract

A global test of goodness-of-fit for the conditional distribution function. We propose a global test of goodness-of-fit to assess the validity of an entertained statistical model by testing simultaneously all the assumptions made about it. This test is based on a local polynomial estimator of the conditional distribution function and on the standard paradigm relating the distance between the nonparametric estimator and the theoretical parametric model. We derive the asymptotic distribution of the resulting test statistic under both the null hypothesis and local alternatives. ***To cite this article: S. Ferrigno, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 341 (2005).***

1. Introduction

Soient X et Y deux variables aléatoires réelles. Dans le but d'expliquer le comportement de Y à partir de celui de X, il est courant de poser un modèle statistique liant le comportement aléatoire de ces variables. Il im-

Adresses e-mail : ferrigno@math.univ-montp2.fr (S. Ferrigno), ducharme@math.univ-montp2.fr (G.R. Ducharme).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.07.003

porte évidemment de s'assurer de la validité de ce modèle pour pouvoir l'utiliser par la suite avec un certain degré de confiance. Pour ce faire, de nombreux tests ont été proposés dans la littérature. Cependant, ils sont tous « directionnels » dans le sens où ils ne testent qu'un seul aspect à la fois du modèle.

Nous proposons un test « global » qui permet de valider d'un seul coup toute la structure prédictive du modèle considéré en s'appuyant sur la fonction de répartition conditionnelle $F(y|x) = P(Y \leqslant y|X = x)$ qui capte globalement toute cette information. Si le modèle posé est exact, cette fonction de répartition prend la valeur $F_0(y|x)$ (que l'on suppose ici entièrement connue) et le problème de valider notre modèle de départ se ramène à celui de tester l'hypothèse nulle

$$H_0 : F(y|x) = F_0(y|x) \quad \forall (x, y). \tag{1}$$

Notre approche s'appuie sur l'idée de comparer un estimateur non paramétrique de $F(y|x)$ à sa valeur supposée $F_0(y|x)$ et à rejeter l'hypothèse nulle si la distance entre ces deux quantités dépasse un seuil critique. Cette idée a été suivie par de nombreux auteurs pour le cas de la fonction de répartition non conditionnelle (Babu et Rao [1] et Neuhaus [6]). L'originalité ici est de s'appuyer sur l'estimateur polynomial local de la fonction de répartition conditionnelle, étudié par Ducharme et Mint El Mouvid [2].

2. L'estimateur polynomial local de $F(y|x)$ et le test

Soient $(X_i, Y_i)_{i=1,\ldots,n}$ des copies indépendantes de (X, Y). Considérons $K(\cdot)$, un noyau positif et $h = h(n)$, une largeur de fenêtre. Si on pose $X_{n\times(p+1)} = ((X_i - x)^j)_{1\leqslant i\leqslant n,\ 0\leqslant j\leqslant p}$ et $P_{n\times n} = \operatorname{diag}(K(\frac{X_i - x}{h}))$, l'estimateur polynomial local de $F(y|x)$ est

$$\hat{F}_n(y|x) = \sum_{i=1}^{n} W^n\left(\frac{X_i - x}{h}\right) I_{\{Y_i \leqslant y\}},$$

où $W^n(t) = e_0^{\mathrm{T}}(X^{\mathrm{T}} P X)^{-1}(1, ht, \ldots, (ht)^p)^{\mathrm{T}} K(t)$ avec $e_0 = (1, 0, \ldots, 0)^{\mathrm{T}}$. Dans la suite, on utilise p impair. Les autres hypothèses considérées sont les suivantes :

(H.1) Le noyau $K(\cdot)$ est une densité de probabilité, symétrique autour de 0, bornée, dérivable, de dérivée bornée et à support compact $[-1, 1]$.
(H.2) La variable aléatoire X a le support compact $[c, d]$.
(H.3) La densité marginale $f(\cdot)$ de X satisfait $0 < m < f(\cdot) < \infty$ pour tout $x \in [c, d]$. Elle est aussi continûment dérivable dans un voisinage de tout $x \in]c, d[$ et de dérivée bornée.
(H.4) Pour tout $x \in [c, d]$, $F_0(y|x)$ est continûment dérivable en y, de dérivée $f_0(y|x)$.
(H.5) Pour tout $y \in \mathbb{R}$, $F_0(y|x)$ est $p + 1$ fois continûment dérivable en x dans un voisinage $V(x)$ de chacun des points de $]c, d[$. De plus, pour tout $u \in V(x)$ avec $x \in]c, d[$, $\sup_{u\in V(x)} \sup_{y\in\mathbb{R}} |F_0^{(p+1)}(y|u)| \leqslant M$.
(H.6) On suppose que $h \to 0$, $nh \to +\infty$, $\frac{\sqrt{\log n}}{\sqrt{nh^2}} \to 0$ et $nh^{2p+5/2} \to 0$.

Le noyau $W^n(\cdot)$ est difficile à manipuler mathématiquement. On en donne une approximation plus stable. Soit $K^\star(t) = e_0^{\mathrm{T}} S^{-1}(1, t, \ldots, t^p)^{\mathrm{T}} K(t)$ où $S_{(p+1)\times(p+1)}$ est la matrice dont les éléments sont les $(\int_{-1}^{1} u^{i+j} K(u)\,\mathrm{d}u)_{0\leqslant i,j\leqslant p}$.

Lemme 2.1. *Soit* $[a, b] \in]c, d[$. *Alors, sous les hypothèses* (H.1)–(H.3) *et* (H.6)*, on a presque sûrement* :

$$\sup_{t\in[-1,1]} \sup_{x\in[a,b]} \left| nh W^n(t) - \frac{K^\star(t)}{f(x)} \right| = \mathrm{O}(h).$$

Démonstration. Ce résultat est une amélioration du Lemme 2.1 de Huang et Fan [4]. On le démontre en étant plus soigneux dans la gestion des restes (voir Ferrigno [3]). □

Pour construire le test de (1), on utilise une distance de type Cramér–von Mises généralisée. L'utilisation obligée du Lemme 2.1 nous amène à restreindre l'hypothèse nulle (1) à $x \in [a,b] \in]c,d[$. On travaille alors avec la statistique de test :

$$T = n\sqrt{h}\int_a^b\int_{\mathbb{R}} \big(\hat{F}_n(y|x) - F_0(y|x)\big)^2 \lambda_0(x,y)\,\mathrm{d}y\,\mathrm{d}x,$$

où $\lambda_0(x,y) = w(x)w_x(y)f_0(y|x)$ avec $w(x)$ et $w_x(y)$ des fonctions de poids supposées positives et bornées.

3. Résultats principaux

Théorème 3.1. *Sous H_0 et sous les hypothèses* (H.1)–(H.6)*, on a* :

$$T - h^{-1/2}a_0 \xrightarrow{\mathcal{L}} N(0,\sigma_0^2),$$

avec

$$a_0 = K^{\star(2)}(0)\int_a^b\int_{\mathbb{R}} \frac{F_0(y|x)(1-F_0(y|x))}{f(x)}\lambda_0(x,y)\,\mathrm{d}y\,\mathrm{d}x, \tag{2}$$

$$\sigma_0^2 = 2K^{\star(4)}(0)\int_a^b\int_{\mathbb{R}^2}\left[\frac{F_0(y\wedge y'|x') - F_0(y|x')F_0(y'|x')}{f(x')}\right]^2 \lambda_0(x',y)\lambda_0(x',y')\,\mathrm{d}y\,\mathrm{d}y'\,\mathrm{d}x', \tag{3}$$

où $K^{\star(\nu)}(\cdot)$ désigne la $\nu^{\text{ème}}$ convolution du noyau $K^{\star}(\cdot)$ par lui-même.

Démonstration. Un développement de $F_0(y|x)$ autour de $F_0(y|X_i)$ mène à

$$\hat{F}_n(y|x) - F_0(y|x) = \sum_{i=1}^n W^n\left(\frac{X_i - x}{h}\right)\big(I_{\{Y_i\leqslant y\}} - F_0(y|X_i)\big) + h^{p+1}R_n(x,y),$$

où le reste $R_n(x,y)$ est presque sûrement borné en x et en y par (H.5). La statistique de test appliquée à ce développement se décompose en $T = T^{\star} + T^{\star\star}$, avec

$$T^{\star} = n\sqrt{h}\int_a^b\int_{\mathbb{R}}\left(\sum_{i=1}^n W^n\left(\frac{X_i - x}{h}\right)\big(I_{\{Y_i\leqslant y\}} - F_0(y|X_i)\big)\right)^2 \lambda_0(x,y)\,\mathrm{d}y\,\mathrm{d}x,$$

et où, par (H.5) et le Lemme 2.1, $T^{\star\star} = \mathrm{o}_p(1)$. De plus, en remplaçant $W(\frac{X_i-x}{h})$ par $\widetilde{W}^n(X_i,x,h) = \frac{K^{\star}((X_i-x)/h)}{nhf(x)}$, $T^{\star} = T_1 + T_2 + \mathrm{o}_p(1)$ avec

$$T_1 = n\sqrt{h}\sum_{i=1}^n\int_a^b\int_{\mathbb{R}} \widetilde{W}^n(X_i,x,h)^2\big(I_{\{Y_i\leqslant y\}} - F_0(y|X_i)\big)^2\lambda_0(x,y)\,\mathrm{d}y\,\mathrm{d}x,$$

$$T_2 = n\sqrt{h}\sum_{1\leqslant i\neq j\leqslant n}\int_a^b\int_{\mathbb{R}} \widetilde{W}^n(X_i,x,h)\widetilde{W}^n(X_j,x,h)\big(I_{\{Y_i\leqslant y\}} - F_0(y|X_i)\big)$$
$$\times\big(I_{\{Y_j\leqslant y\}} - F_0(y|X_j)\big)\lambda_0(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

Or, $T_1 - h^{-1/2}a_0 \xrightarrow{P} 0$, où a_0 est donnée en (2). La statistique T_2 est, à une constante près, une U-statistique dégénérée. Sa normalité asymptotique est obtenue en vérifiant les conditions du Théorème 2.1 de de Jong [5]. On trouve alors $T_2 \xrightarrow{\mathcal{L}} N(0, \sigma_0^2)$, où σ_0^2 est donnée en (3). □

On affine l'analyse du test en étudiant sa puissance asymptotique locale. Considérons la suite d'alternatives locales

$$H_{1n} : F(y|x) = F_{1n}(y|x) = F_0(y|x) + (n\sqrt{h})^{-1/2} G(y|x),$$

où la fonction $G(y|x)$ est $p+1$ fois dérivable en x et de dérivée $(p+1)^{\text{ème}}$ bornée. On suppose aussi pour simplifier que la densité marginale de X est la même sous H_0 comme sous H_{1n}.

Théorème 3.2. *Sous H_{1n} et sous les hypothèses du Théorème* 3.1*, on a* :

$$T - h^{-1/2}a_0 \xrightarrow{\mathcal{L}} N\big(b_0, \sigma_0^2\big),$$

où a_0 et σ_0^2 sont données respectivement en (2) *et* (3) *et où*

$$b_0 = \int_a^b \int_{\mathbb{R}} G^2(y|x)\lambda_0(x, y)\,\mathrm{d}y\,\mathrm{d}x.$$

Démonstration. Par une décomposition de la statistique de test sous H_{1n}, on a :

$$T = T_1^\star + T_2^\star + T_3^\star,$$

avec

$$T_1^\star = n\sqrt{h} \int_a^b \int_{\mathbb{R}} \big[\hat{F}_n(y|x) - F_{1n}(y|x)\big]^2 w(x) w_x(y) F_{1n}(\mathrm{d}y|x)\,\mathrm{d}x,$$

$$T_2^\star = \int_a^b \int_{\mathbb{R}} G^2(y|x) w(x) w_x(y) F_{1n}(\mathrm{d}y|x)\,\mathrm{d}x,$$

et où, par l'inégalité de Bienaymé–Tchebytchev, $T_3^\star = \mathrm{o}_p(1)$. De plus, $T_2^\star = b_0 + \mathrm{o}(1)$ et la normalité asymptotique de $T_1^\star$ est obtenue par une version triangulaire du Théorème 2.1 de de Jong [5]. □

Notons que ces résultats peuvent être généralisés au contexte où la fonction à tester sous H_0 dépend de paramètres inconnus (voir Ferrigno [3]). Notons par ailleurs qu'il est aussi possible de généraliser la procédure de test au cas où X est multidimensionnel.

Références

[1] J. Babu, C.R. Rao, Goodness-of-fit tests when parameters are estimated, Sankhyā 66 (2004) 63–74.
[2] G.R. Ducharme, M. Mint El Mouvid, Convergence presque sûre de l'estimateur linéaire local de la fonction de répartition conditionnelle, C. R. Acad. Sci. Paris, Sér. I 333 (2001) 873–876.
[3] S. Ferrigno, Un test d'adéquation global pour la fonction de répartition conditionnelle, Ph.D. thesis, Université Montpellier II, 2004.
[4] L.-S. Huang, J. Fan, Nonparametric estimation of quadratic regression functionals, Bernoulli 5 (5) (1999) 927–949.
[5] P. de Jong, A central limit theorem for generalized quadratic forms, Probab. Theory Related Fields 75 (1987) 261–277.
[6] G. Neuhaus, Asymptotic theory of goodness-of-fit when parameters are present: a survey, Math. Operationsforsch. Statist. Ser. Statist. 10 (1979) 479–494.