Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 338 (2004) 317–320
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Test d'adéquation en régression non-linéaire, cas des fonctions monotones

Bénédicte Fontez [a], Gilles R. Ducharme [b]

[a] *Institut de l'élevage, parc scientifique agropolis 34397 Montpellier cedex 5, France*
[b] *Laboratoire de probabilités et statistique, cc51, Université Montpellier II, place Eugène Bataillon 34095, Montpellier cedex 5, France*

Reçu le 25 juin 2003 ; accepté après révision le 9 décembre 2003

Présenté par Paul Deheuvels

Résumé

Nous proposons un test d'adéquation pour des modèles de régression non-linéaire monotone. Ce test s'appuie sur le paradigme des tests lisses de Neyman (1937) pour tester l'adéquation des densités. Il s'applique tout particulièrement aux fonctions de croissance. Nous complétons cette approche en incorporant une méthode de sélection automatique des paramètres du test par les données. ***Pour citer cet article : B. Fontez, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

Abstract

Goodness-of-fit test for monotonous nonlinear regression models. We propose a goodness-of-fit test for monotonous nonlinear regression models. The test is based on an adaptation, to the regression context, of the smooth test paradigm of Neyman (1937) for testing the distribution of a sample. We complete his approach with a data-driven criterion for the test's parameters. ***To cite this article: B. Fontez, G.R. Ducharme, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

1. Introduction

Pour étudier l'évolution d'un phénomène dans le temps ou la relation entre deux variables, la modélisation paramétrique reste auprès des praticiens l'approche privilégiée car elle permet de résumer cette relation par une courbe, $F(x;\beta)$, dépendant des paramètres β. Ces derniers sont des outils d'interprétation importants car, entre autres, on peut tester leur valeur pour inférer des hypothèses. Mais avant, le modèle $F(x;\beta)$ doit être validé. Or, ce modèle est souvent non linéaire et dans ce contexte, l'utilisation des outils classiques tels que le R^2 est controversée (Ratkowsky [13]). Ainsi, les méthodes les plus populaires pour construire un test d'adéquation utilisent une estimation non paramétrique $\widehat{G}(\cdot)$ de la relation et la comparent à un estimateur paramétrique de $F(\cdot;\beta)$. Sur ce registre, citons les travaux de Kuchibathla et Hart [9], Antoniadis, Gijbels et Grégoire [1]. Une autre approche consiste à adapter les outils existant pour tester l'adéquation de densités au contexte de la régression.

Adresses e-mail : benedicte.fontez@inst-elevage.asso.fr (B. Fontez), ducharme@math.univ-montp2.fr (G.R. Ducharme).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2003.07.008

Dans cette optique, on peut citer les tests proposés notamment par Stute [14], Diebolt et Zuber [2] et Harel [8], dont la démarche correspond à l'esprit d'un test de type Neyman [12]. L'approche que nous présentons ici combine les deux démarches. Nous proposons un test de type Neyman où l'approximation $\widehat{G}(\cdot)$ est une combinaison linéaire de K fonctions d'une base orthonormale, avec K déterminé à partir des données. Par construction, notre $\widehat{G}(\cdot)$ conserve la monotonie de la courbe $F(\cdot;\beta)$. Ainsi, nous proposons un test ciblé qui, dans le contexte des courbes de croissance, montre de bonnes puissances lors des simulations. Nous explorons, en plus, une généralisation au cas d'un faisceau de courbes où on observe la croissance d'une espèce sur plusieurs individus.

2. Résultats

Nous supposons que le phénomène de croissance Y, observé dans le temps x, est continu dérivable et monotone et qu'il est modélisé séparément pour N individus. On pose ainsi le modèle :

$$Y_{it} = G_i(x_{it}) + \varepsilon_{it}, \quad t = 1, \ldots, n_i,\ i = 1, \ldots, N. \tag{1}$$

Dans cette expression, les temps d'observations x_{it} sont fixés et les erreurs ε_{it} sont supposées iid, de moyenne nulle et de variance finie σ_i^2. Ce contexte peut être étendu au cas où la structure de la matrice de variance-covariance, Σ_i, de $(\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ est connue ou peut être estimée. On suppose que $G_i(\cdot)$ prend la forme fonctionnelle $F(\cdot;\beta_i)$, supposée continue dérivable ; on souhaite donc tester l'hypothèse :

$$H_0 : G_i(\cdot) \in \left\{F(\cdot;\beta_i),\ \beta_i \in \Phi \subseteq R^P\right\} \quad \forall i = 1, \ldots, N, \tag{2}$$

contre l'alternative H_1 qu'au moins un des $G_i(\cdot)$ n'appartient pas à cette famille. On utilise l'estimateur des moindres carrés pour estimer le vecteur des paramètres β, de dimension P et $\hat{\sigma}_i^2 = \sum_t \hat{\varepsilon}_{it}^2/(n_i - P)$ pour estimer σ_i^2 où $\hat{\varepsilon}_{it}$ désigne le résidu au temps x_{it} de l'individu i sous H_0. On suppose enfin que la croissance est observée sur un intervalle de temps fini $[a, b]$ où, pour simplifier les notations, on pose $\min\{x_{it},\ t \geqslant 1\} > a$.

Proposition 2.1. *Statistique du test pour $N = 1$: on note $\{\pi_i\}_{i\geqslant 0}$ la base des polynômes de Legendre orthonormés sur $L^2[0,1]$, $\widehat{Z}_t = (F(x_t;\hat{\beta}) - F(a;\hat{\beta}))/(F(b;\hat{\beta}) - F(a;\hat{\beta}))$ la variable transformée sur $[0,1]$. Dénotons $\widehat{\mathbf{T}}_{n\times K}$ et $\widehat{\mathbf{V}}_{n\times P}$ les matrices de termes respectifs, $\int_0^{\widehat{Z}_t} \pi_k(z)\,\mathrm{d}z$ et $\partial F(x_t;\hat{\beta})/\partial\beta_p$ où $p = 1, \ldots, P$, $k = 1, \ldots, K$ et $t = 1, \ldots, n$. La statistique de test*

$$\widehat{R}_K = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}^{\mathrm{T}}\widehat{\mathbf{T}}\big(\widehat{\mathbf{T}}^{\mathrm{T}}\widehat{\mathbf{T}} - \widehat{\mathbf{T}}^{\mathrm{T}}\widehat{\mathbf{V}}\big(\widehat{\mathbf{V}}^{\mathrm{T}}\widehat{\mathbf{V}}\big)^{-1}\widehat{\mathbf{V}}^{\mathrm{T}}\widehat{\mathbf{T}}\big)^{-1}\widehat{\mathbf{T}}^{\mathrm{T}}\hat{\varepsilon} \tag{3}$$

converge en loi sous H_0 vers un χ_K^2.

Dans le cas où la matrice de variance-covariance du vecteur des erreurs (Σ) est connue ou estimable selon les méthodes décrites par exemple dans Gallant [7], le test s'applique moyennant quelques modifications. On utilise l'estimateur des moindres carrés pondérés de β donné par Gallant [7] et on considère la statistique de test :

$$\widehat{R}_K^* = \frac{1}{\hat{\sigma}^2}\hat{\varepsilon}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\mathbf{T}}\big(\widehat{\mathbf{T}}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\mathbf{T}} - \widehat{\mathbf{T}}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\mathbf{V}}\big(\widehat{\mathbf{V}}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\mathbf{V}}\big)^{-1}\widehat{\mathbf{V}}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{\mathbf{T}}\big)\widehat{\mathbf{T}}^{\mathrm{T}}\widehat{\Sigma}^{-1}\hat{\varepsilon},$$

où $\widehat{\Sigma}$ est une estimation $\sqrt{n}$-consistente de Σ et $\hat{\sigma}^2 = \hat{\varepsilon}^{\mathrm{T}}\widehat{\Sigma}^{-1}\hat{\varepsilon}/n$.

Proposition 2.2. *Choix de K adapté aux données : on choisit deux entiers $1 \leqslant d \leqslant D$ et on définit $\widehat{K} = \min[\mathrm{Arg\,max}_{d\leqslant s\leqslant D}\{\widehat{R}_s - s\log(n)\}]$. La statistique $\widehat{R}_{\widehat{K}}$ converge en loi vers un χ_d^2 sous H_0.*

La loi asymptotique n'est pas toujours une bonne approximation de la loi de $\widehat{R}_{\widehat{K}}$ pour de petits échantillons. On peut calculer des quantiles plus précis en résolvant numériquement l'approximation suivante :

$$P\left(\widehat{R}_{\widehat{K}} \leqslant x\right) \approx \Delta_1(\log n)\left[\int_0^{x-\log n} \Delta_1(x-z)\delta_d(z)\,\mathrm{d}z - \Delta_1(\log n)\Delta_d(x-\log n)\right] + \Delta_d(x)\int_0^{\log n} \Delta_1(2\log n - z)\delta_1(z)\,\mathrm{d}z + \int_{\log n}^{x-\log n}\int_{\log n}^{x-y} \Delta_d(x-y-z)\delta_1(y)\delta_1(z)\,\mathrm{d}z\,\mathrm{d}y + \int_0^{\log n}\int_{2\log n-y}^{x-y} \Delta_d(x-y-z)\delta_1(y)\delta_1(z)\,\mathrm{d}z\,\mathrm{d}y, \tag{4}$$

où Δ_d et δ_d désignent respectivement la fonction de répartition d'un χ^2_d et sa densité.

Proposition 2.3. *Généralisation à $N > 1$: on note $\widehat{R}^i_K$ la statistique de la Proposition* 2.1 *pour l'individu i. Si $\forall i \leqslant N,\ n_i \to \infty$ tel que $n_i / \sum_i n_i \to \lambda_i$ où $0 < \lambda_i < \infty$, alors la statistique de test $\widehat{R}_{N,K} = \sum_{i=1}^N \widehat{R}^i_K$ converge en loi sous H_0 vers un χ^2_{NK}. Soit $\widehat{K} = \min \operatorname{Arg}\max_{d\leqslant s\leqslant D}\{\widehat{R}_{N,s} - Ns\log(n)\}$, la statistique $\widehat{R}_{N,\widehat{K}}$ converge en loi sous H_0 vers un χ^2_{Nd}.*

3. Démarche

Nous allons expliciter la démarche pour $N = 1$ individu, sachant que le cas $N > 1$ s'obtient facilement en appliquant la même démarche aux N individus. Le paradigme des tests lisses de type Neyman [12] propose de construire une famille de fonctions emboîtant $F(x;\beta)$, notée $\mathcal{G}_K = \{G(x;\beta;\theta),\ \beta \in \Phi \subseteq R^P,\ \theta \in \Theta \subseteq R^K\}$, telle que $G(x;\beta;0) = F(x;\beta)$, puis d'utiliser le test des scores (test de Rao ou du Multiplicateur de Lagrange) pour tester la nullité du vecteur de paramètres θ. Dans le cas des courbes de croissance, cette approche permet de définir comme alternative une fonction de croissance $G(x;\beta;\theta)$, concentrant ainsi la puissance sur l'ensemble des fonctions croissantes. Quand la courbe est décroissante, ce test reste légitime si l'alternative conserve la même propriété de monotonie.

La famille $\mathcal{G}_K$ est construite en utilisant la distance de Hellinger qui sépare les pseudo densités tronquées à $[a,b]$ $f^*(x;\beta)$ de $F^*(x;\beta) = (F(x;\beta) - F(a;\beta))/(F(b;\beta) - F(a;\beta))$ et $g^*(x)$ de $G^*(x) = (G(x) - G(a))/(G(b) - G(a))$. La fonction $\sqrt{g^*/f^*}$ existant presque partout par rapport à $f^*\,\mathrm{d}x$, on la développe sur $L^2(f^*)$, l'espace des fonctions de carré intégrables muni du produit scalaire $(f_1, f_2)_{f^*} = \int_a^b f_1(t)f_2(t)f^*(t,\beta)\,\mathrm{d}t$, selon une base orthonormale $\{h_i\}_{i\geqslant 0}$. On obtient $g^*(x) = f^*(x;\beta)(\sum_{k=0}^\infty \theta^*_k h_k(x;\beta))^2$, avec $\theta^*_k = (\sqrt{g^*/f^*}, h_k)_{f^*}$ et $\sum_{k=0}^\infty (\theta^*)^2 \equiv 1$. On pose $\theta_k = \theta^*_k/\theta^*_0 (\theta^*_0 \neq 0)$, de sorte que $g^*(x) = f^*(x;\beta)(1 + \sum_{k=1}^\infty \theta_k h_k(x;\beta))^2(1 + \sum_{k=1}^\infty \theta^2)^{-1}$. Généralement $|\theta_k|$ décroît rapidement vers 0 quand $k \to \infty$ et le développement peut raisonnablement être tronqué à l'ordre K. En intégrant cette approximation, on construit la famille d'emboîtement :

$$G(x;\beta;\theta) = \int_a^x f(t;\beta)\frac{(1 + \sum_{k=1}^K \theta_i h_i(t;\beta))^2}{1 + \sum_{k=1}^K \theta_k^2}\,\mathrm{d}t + F(a;\beta), \tag{5}$$

où $f(x;\beta)$ est la dérivée de $F(x;\beta)$. Pour plus d'information sur la construction de ce type de famille d'emboîtement, on pourra se reporter à l'article de Ducharme [3]. Plusieurs bases $\{h_i\}$ sont possible pour (5). Dans le présent contexte, il est commode de suivre les traces de Neyman [12] et transformer les données pour utiliser toujours la base des polynômes de Legendre.

Si K est tel que la famille $\mathcal{G}_K$ approche suffisamment la vraie fonction de croissance $G(\cdot)$; l'hypothèse H_0 de (2) se ramène à tester $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Dans ce contexte, le test des scores possède de bonnes

propriétés de puissance (Gallant [7]) et ne nécessite pas l'estimation des paramètres du modèle sous H_1. Sous les hypothèses de Gallant [7], pp. 255–256), nous pouvons appliquer les théorèmes de son chapitre 3 qui montrent la convergence de $\widehat{R}_K$ vers la loi d'un χ^2_K sous H_0 et d'un $\chi^2_K(\alpha)$ décentré sous l'hypothèse alternative contigue où on laisse tendre H_{1n} vers H_0 en posant $\theta_n = \theta^*/\sqrt{n}$ et $\beta = \beta_0$ dans l'Éq. (5). Dans ce contexte, l'espérance de l'estimateur des moindres carrés vérifie $E(\hat{\beta}_n) = \beta_0 + \beta^*/\sqrt{n} + o(1/\sqrt{n})$. Aussi, pour θ^* et β^* des vecteurs de dimension K et P finis et fixés, on note $\Delta = (\theta^*, \beta^*)$. Avec les notations supplémentaires $\Lambda = (\theta, \beta)$ et $\Omega = \lim_{n\to\infty}(n\sigma^2)^{-1}(\partial G(x;\beta_0;0)/\partial\Lambda)^{\mathrm{T}}(\partial G(x;\beta_0;0)/\partial\Lambda)$, le paramètre de décentrage α vaut $\Delta^{\mathrm{T}}\Omega\Delta$. Pour plus de détail, on peut se référer aux démonstrations complètes dans la thèse de Fontez [6].

Nous avons pour l'instant supposé que K était adéquatement choisi. Neyman [12] proposait de prendre $K \in \{2, 3, 4\}$ dans son article, mais il est possible d'adapter le choix de K aux données. Citons les travaux de Ledwina [10] sur le critère de Schwarz, Antoniadis Gijbels et Grégoire [1] sur le LIC, Kuchibathla et Hart [9] et Fontez [6] sur le C_p. Plus récemment, Lee et Hart [11] ou Fan [5] entre autres ont essayé des méthodes de seuillage. Nous avons choisi le critère de Schwarz car il est simple à appliquer (cf. Proposition 2.2) et des quantiles ou des approximations précises pour de petits échantillons peuvent être calculés, cf. Ducharme et Lafaye de Michaux [4]. Dans notre cas, nous avons affiné, à l'ordre supérieur, leur correction pour l'approximation des quantiles, la formule est donnée en (4).

Références

[1] A. Antoniadis, I. Gijbels, G. Grégoire, Model selection using wavelet decomposition and applications, Biometrika 84 (1997) 751–763.
[2] J. Diebolt, J. Zuber, Goodness of fit tests for nonlinear heteroscedastic regression models, Statist. Probab. Lett. 42 (1999) 53–60.
[3] G. Ducharme, Goodness of fit tests for the inverse gaussian and related distributions, Test 10 (2001) 271–290.
[4] G. Ducharme, E. Lafaye de Michaux, Goodness of fit tests of normality for the innovation in ARMA models, J. Time Series (2003) sous-presse.
[5] J. Fan, Test of significance based on wavelet thresholding and Neyman's truncation, JASA 91 (1996) 674–688.
[6] B. Fontez, Test d'adéquation des résidus pour la régression non-linéaire, application aux courbes de croissance en foresterie, Ph.D. thesis, Université Montpellier II, 2001.
[7] A. Gallant, Nonlinear Statistical Models, Wiley, New York, 1987.
[8] M. Harel, Une méthode semi-paramétrique pour tester un modèle de régression, C. R. Acad. Sci. Paris, Ser. I 336 (2003) 601–604.
[9] M. Kuchibhatla, J. Hart, Smoothing-based lack of fit tests: variations on a theme, Nonparametric Statist. 7 (1996) 1–22.
[10] T. Ledwina, Data-driven version of neyman's smooth test of fit, JASA 89 (1994) 1000–1005.
[11] G. Lee, J. Hart, An l_2 error test with order selection and thresholding, Statist. Probab. Lett. 39 (1998) 61–72.
[12] J. Neyman, Smooth test for goodness of fit, Skand. Aktuar. 20 (1937) 149–199.
[13] D. Ratkowsky, Handbook of Nonlinear Regression Models, Dekker, New York, 1989.
[14] W. Stute, Nonparametric model checks for regression, Ann. Statist. 25 (1997) 613–641.