Test de linéarité dans un modèle de régression non paramétrique

Zaher Mohdeb

TEST DE LINEARITE DANS UN MODELE DE REGRESSION NON PARAMETRIQUE

Zaher MOHDEB
Département de Mathématiques
Université Mentouri, Constantine, Algérie
E-mail: z.mohdeb@gmail.com

Résumé: Une procédure de test d'hypothèse linéaire sur la fonction de régression f dans un modèle de régression non paramétrique est proposée. Plus précisément, on teste l'hypothèse que f est un élément de E, où E est un espace vectoriel de dimension finie. En supposant que les fonctions considérées sont höldériennes d'ordre plus grand que $1/2$ et on obtient le comportement asymptotique du test proposé, on a donc ainsi le niveau et la puissance asymptotique du test. Une étude par simulation a été menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposée.

Abstract: A procedure for testing linear hypothesis on the regression function f in a nonparametric regression model. More precisely, we test that f is an element of E, where E is a finite dimensional vector space. We assume that the functions satisfy the Hölder condition with order strictly greater than $1/2$, and we obtain the asymptotic weak behaviour of the proposed test, then we have the level and the asymptotic power of the test. A simulation study is conducted, for small sample size, to demonstrate the performance the proposed test.

Mots clés: Régression non paramétrique, test de linéarité.

1. Introduction

On considère le modèle de régression suivant

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\ , \quad i = 1, \ldots, n\,, \tag{1}$$

où f est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{1,n} = 0 < t_{2,n} < \cdots < t_{n,n} = 1$, est un échantillonnage fixé de l'intervalle $[0,1]$. Les erreurs $\varepsilon_{i,n}$ forment un tableau triangulaire de variables aléatoires d'espérance nulle et variance finie σ^2, tel que pour tout n les variables aléatoires $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ soient indépendantes. Notre objectif est la construction de tests d'hypothèses linéaires sur la fonction de régression f. Plus précisément, soient $g_1(t), \ldots, g_p(t)$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit E_p l'espace vectoriel engendré par $g_1, \ldots, g_p$. On veut tester l'hypothèse:

$$\mathrm{H}_0 : \quad f \in E_p \qquad \text{contre} \qquad \mathrm{H}_1 : \quad f \notin E_p\,. \tag{2}$$

Le problème de test d'hypothèses dans le modèle (1) a donné lieu à de nombreux travaux, l'usage des coefficients de Fourier empiriques de f est abordé par Eubank et Spiegelman (1990) et Mohdeb et Mokkadem (2001). Dette et Munk (1998) introduit une statistique de test basée sur l'estimation du carré de la distance de f à E_p. Mohdeb et Mokkadem (2002) proposent une statistique de test basée sur une approche similaire à celle de Dette et Munk (1998), mais sans utilisation de poids et montrent qu'elle a le même comportement asymptotique. Mohdeb et Mokkadem (2004 a, 2004 b) proposent également une autre statistique de test basée sur une autre estimation du carré de la distance de f à E_p .
Dans ce travail, on procède comme dans Mohdeb et Mokkadem (2002), mais en utilisant une autre estimation de la variance des erreurs $(\varepsilon_{i,n})$ intervenant dans la statistique de test. Nous obtenons une loi asymptotique de celle-ci avec une variance différente de celle donnée dans Mohdeb et Mokkadem (2002). Des simulations ont été menées pour étudier la performance du test proposé.

2. Résultat principal

Soit h une fonction densité sur l'intervalle $[0,1]$, positive et de Hölder d'ordre $\gamma > 1/2$ et que l'échantillonnage $\{t_{1,n},\ldots,t_{n,n}\}$ est associé à h. On note $L^2(d\mu)$ où $d\mu = h(t)dt$, l'espace des fonctions de carré intégrables, muni du produit scalaire usuel $\langle .,. \rangle$. La distance entre f et E_p est notée par $D(f)$. On suppose que les hypothèses suivantes sont satisfaites.

(C1) $\max\limits_{i=2,\ldots,n} \left| \int_{t_{i-1,n}}^{t_{i,n}} h(t)\,dt - \dfrac{1}{n} \right| = o\left(\dfrac{1}{n}\right)$;

(C2) h , f , x_k, , $k = 1,\ldots,p$, sont des fonctions de Hölder d'ordre $\gamma > 1/2$;

(C3) $\forall n$, $\varepsilon_{1,n},\ldots,\varepsilon_{n,n}$ sont indépendantes et $\exists C \in \mathbb{R}^+$ tel que $E(\varepsilon_{i,n}^4) < C$, $\forall i, n$.

Soit $(\widehat{\beta}_1,\ldots,\widehat{\beta}_p)' = \underset{(b_1,\ldots,b_p)\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \dfrac{1}{n}\sum_{i=1}^{n} \left| Y_{i,n} - \sum_{k=1}^{p} b_k f_k(t_{i,n}) \right|^2 \right\}$.
L'approche que nous proposons repose sur le faite que l'hypothèse:
"$f \in E_p$" est équivalente à "$\phi = f - \sum_{k=1}^{p} \widehat{\beta}_k g_k \in E_p$".
Cela nous amène à utiliser la distance $D(\phi)$ entre ϕ et E_p au lieu de $D(f)$.

On suppose que $E_p \subset L^2(d\mu)$ et on veut estimer:

$$D^2(\phi) = \min_{u\in E_p} \int_0^1 |\phi(t) - u(t)|^2 h(t)dt\,.$$

Cette distance peut aussi s'écrire sous la forme:

$$D^2(\phi) = \frac{\Gamma(\phi, g_1, \ldots, g_p)}{\Gamma(g_1, \ldots, g_p)}\,, \tag{3}$$

où $\Gamma(u_1, \ldots, u_p)$ est le déterminant de Gram $|(\langle u_i, u_j)|_{i,j=1,\ldots,k}$ pour $u_1, \ldots, u_k$ dans $L^2(d\mu)$.
Pour estimer $D^2(\phi)$, nous introduisons les observations $X = (X_{1,n}, \ldots, X_{n,n})$ où $X_{i,n} = Y_{i,n} - \sum_{k=1}^{p} \widehat{\beta}_k g_k(t_{i,n})\,, \quad i = 1, \ldots, n\,.$ On procède comme dans Dette et Munk (1998), mais sans utilisation de poids, pour construire un estimateur empirique de $D^2(\phi)$. Soit $\Delta_{i,n} = t_{i,n} - t_{i-1,n}$, $i = 2, \ldots, n$, $\Delta_{1,n} = \Delta_{2,n}$, $W = \mathrm{diag}\Big(\Delta_{i,n} h(t_{i,n})\Big)_{i=1,\ldots,n}$ et $g_{k,n} = \Big(g_k(t_1), \ldots, g_k(t_n)\Big)'$, $k = 1, \ldots, n$.
Soit $E_{p,n}$ le sous espace de $\mathbb{R}^n$ engendré par $(g_{1,n}, \ldots, g_{p,n})$ et $\Pi_n^{\perp}$ la matrice de projection sur l'hortogonal à $E_{p,n}$.
On définit $\Gamma_n(X, g_1, \ldots, g_p)$ comme étant le déterminant obtenu en remplaçant dans $\Gamma(\phi, g_1, \ldots, g_p)$, les produits scalaires $\langle \phi, \phi \rangle$ et $\langle \phi, g_k \rangle$, $k = 1, \ldots, p$, respectivement par les quantités: $X'WX = \sum_{i=1}^{n} \Delta_{i,n} h(t_{i,n}) X_{i,n}^2$ et $X'W g_{k,n} = \sum_{i=1}^{n} \Delta_{i,n} h(t_{i,n}) g_k(t_{i,n}) X_{i,n}\,, \quad k = 1, \ldots, p\,.$ On obtient un estimateur de $D^2(\phi)$ donné par:

$$\widehat{D}_n^2 = \frac{\Gamma_n(X, g_1, \ldots, g_p)}{\Gamma(g_1, \ldots, g_p)} - \widehat{\sigma}_R^2 \mathrm{tr}(W \Pi_n^{\perp})\,,$$

où $\widehat{\sigma}_R^2$ est l'estimateur de Rice (1984), défini par $\widehat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_{i,n} - Y_{i-1,n})^2\,.$
On rejette H_0 si $\widehat{D}_n^2 > c$. Le seuil ainsi que la puissance asymptotique du test sont obtenus en étudiant la loi asymptotique de $\widehat{D}_n^2$. On a le comportement asymptotique de la statistique de test proposée.

Théorème 1 *Si les hypothèses* $(C1)$, $(C2)$ *et* $(C3)$ *sont satisfaites alors*

$$\sqrt{n}\left(\widehat{D}_n^2 - D(f)^2\right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^4 + 4\sigma^2 D^2(f)\right)\,, \quad \textit{quand} \quad n \to \infty\,.$$

Proposition 1 *Considérons les alternatives locales de la forme* $f(t) = u(t) + c_n v(t)$, *où* u *est une fonction du sous-espace* U_p, v *est une fonction orthogonale à* U_p *et* c_n *est une suite tendant vers zéro; alors*

$$\sqrt{n}\widehat{D}_n^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(\int_0^1 v^2(t) h(t)\, dt,\, \sigma^4\right) \quad \textit{si} \quad c_n = n^{-1/4}\,,$$

et

$$\sqrt{n}\widehat{D}_n^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^4\right) \quad \textit{si} \quad c_n = o(n^{-1/4})\,.$$

Ce qui signifie que le test peut détecter les alternatives locales convergeant vers l'hypothèse nulle avec une vitesse plus petite que $n^{-1/4}$.

Le théorème précédent nous donne le niveau et la puissance du test. La procédure de test est la suivante: soit $\widehat{\sigma}^2$ un estimateur consistant de σ^2, puisque $D^2(f) = 0$, lorsque $f \in E_p$, on rejette l'hypothèse nulle H_0, au niveau α, si $\sqrt{n}\widehat{D}_n^2/\widehat{\sigma}^2 > c_{1-\alpha}$, où $c_{1-\alpha}$ est le $(1-\alpha)$-quantile d'une loi normale standard.

3. Simulations

Dans nos simulations, on étudie le test de l'hypothèse H_0: $f \in E_2$, où E_2 est le sous-espace vectoriel de $L^2(d\mu)$, engendré par $g_1(t) = t$ et $g_2(t) = 1$. On a mené une étude Monte Carlo en simulant le modèle (1), avec $t_{i,n} = (i-1)/(n-1)$, $\varepsilon_{i,n} \sim$ i.i.d.$\mathcal{N}(0, \sigma^2)$. L'étude consiste à faire une comparaison avec la statistique $\widehat{M}_n^2$ proposée par Dette et Munk (1998) pour une petite taille d'échantillon $n = 50$. Pour étudier la puissance du test au niveau $\alpha = 0.05$, comme alternative à l'hypothèse H_0, on considère la fonction suivante: $f(t) = a_1 t + a_2 t + \gamma t e^{-2t}$, avec plusieurs valeurs de γ dans l'intervalle $[0, 2]$. Les résultats obtenus par les simulations montrent que les tests basés respectivement sur les deux statistiques $\widehat{M}_n^2$ et $\widehat{D}_n^2$ sont comparables.

Bibliographie

[1] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 778-800.

[2] Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.* **85**, 410, 387-392.

[3] Mohdeb, Z. and Mokkadem, A. (2001). Testing Hypotheses On Fourier Coefficients in Nonparametric Regression Model. *Journal of Nonparametric Statistics*, **13**, 605-629.

[4] Mohdeb, Z. and Mokkadem, A. (2002). Testing the goodness-of-fit of a linear model in nonparametric regression. *Goodness-of-Fit Tests and Model Validity,* Stat. Ind. Technol., *Birkhauser Boston, Boston, MA.*, 185-193.

[5] Mohdeb, Z. and Mokkadem, A. (2004 a). Average squared residuals approach for testing linear hypotheses in nonparametric regression. The International Conference on Recent Trends and Directions in Nonparametric Statistics. *J. Nonparametr. Stat.* **16**, no. 1-2, 3–12.

[6] Mohdeb, Z. and Mokkadem, A. (2004 b). On the Use of Nonparametric Regression for Testing Linear Hypotheses. *Ann. I.S.U.P.* **48**, no. 3, 63–77.

[7] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Stat.* **12**, 1215-1230.