Tests d'hypothèses dans un modèle de régression non paramétrique, cas non höldérien

Zaher Mohdeb

Laboratoire de Mathématiques et Sciences de la Décision
Département de Mathématiques, Université Constantine 1
Constantine, Algérie
E-mail: zaher.mohdeb@umc.edu.dz

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_p, où E_p est un espace vectoriel de dimension finie. En supposant que les fonctions considérées sont Riemann-intégrables et on obtient le comportement asymptotique du test proposé, on a donc ainsi le niveau et la puissance asymptotique du test.

Mots-clés. Hypothèse linéaire, Régression non paramétrique, Régression non linéaire.

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_p, where E_p is a finite dimensional vector space. We assume that the functions are Riemann-integrable, and we obtain the asymptotic weak behaviour of the proposed test, then we have the level and the asymptotic power of the test.

Keywords. Linear hypothesis, Nonparametric regression, Nonlinear regression.

1 Introduction

L'objet du présent travail est de construire des tests d'hypothèses linéaires dans le modèle de régression non paramétrique sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative. On considère donc 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 $t_{i,n}$, $i = 1, \ldots, n$, 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 de variance finie σ^2.
Soient $x_1(t), \ldots, x_p(t)$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit E_p l'espace vectoriel engendré par $x_1, \ldots, x_p$. On veut tester l'hypothèse:

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

La plupart des travaux sur les tests d'hypothèses dans le modèle (1) supposent des conditions de régularité sur $f,\ \ x_1,\ldots,x_p$; généralement ces fonctions sont supposées höldériennes. On peut citer Eubank et Spiegelman (1990), Eubank et Hart (1992) et Härdle et Mammen (1993). Dette et Munk (1998) ont abordé le test (2) avec l'hypothèse f höldérienne d'ordre $\gamma > 1/2$. Mohdeb et Mokkadem (2004) 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 suppose que $f,\ x_1,\ldots,x_p$ sont Riemann-intégrables; sous cette seule condition sur les fonctions, on établit un théorème de convergence en loi qui permet de construire des tests avec des hypothèses non régulières.

2 Résultat principal

On considère le modèle de régression (1) et E_p est l'espace vectoriel engendré par des fonctions fixées $x_1(t),\ldots,x_p(t)$ définies sur $[0,1]$ et linéairement indépendantes.
Nos hypothèses sont les suivantes:

- (A1) $\max\limits_{i=2,\ldots,n}\left|(t_{i,n}-t_{i-1,n})-\dfrac{1}{n}\right| = o\left(\dfrac{1}{n}\right)$;
- (A2) $\forall n,\ \varepsilon_{1,n},\ldots,\varepsilon_{n,n}$ sont indépendantes et $\exists C\in\mathbb{R}^+$ tel que $E(\varepsilon_{i,n}^4)<C,\ \ \forall i,n$;
- (A3) Les fonctions $f\,,\ \ x_1,\ldots,x_p$ sont Riemann-intégrables.

Les fonctions que nous considérons, sont aussi dans $L^2(dt)$ muni de son produit scalaire usuel

$$\langle u,v\rangle=\int_0^1 u(t)\,v(t)\,dt \quad \text{et} \quad \|u\|_2^2=\int_0^1 u^2(t)\,dt\,, \qquad u,v\in L^2(dt)\,.$$

On pose

$$\mathcal{D}^2(f)=\min_{v\in E_p}\|f-v\|_2^2$$

le carré de la distance de f à E_p.
Pour construire notre statistique de test, on utilise une estimation empirique de $\mathcal{D}^2(f)$. Pour cela, on pose $Y_n=(Y_{1,n},\ldots,Y_{n,n})'$, $\varepsilon_n=(\varepsilon_{1,n},\ldots,\varepsilon_{n,n})'$, $f_n=\Big(f(t_{1,n}),\ldots,f(t_{n,n})\Big)'$, $x_{k,n}=\Big(x_k(t_{1,n}),\ldots,x_k(t_{n,n})\Big)'$, $\ k=1,\ldots,p$, et $X=(x_{1,n},\ldots,x_{p,n})$ c'est-à-dire la matrice dont le $(i,j)^{ieme}$ élément est $x_j(t_i)\,,\ \ i=1,\ldots,n$ et $j=1,\ldots,p$.

On note aussi $E_{p,n}$, le sous-espace de $\mathbb{R}^n$ engendré par $\{x_{1,n},\ldots,x_{p,n}\}$.

Dans $\mathbb{R}^n$, on utilise le produit scalaire usuel et on pose

$\Pi_n=X(X'X)^{-1}X'$, la matrice de projection sur $E_{p,n}$ et
$\Pi_n^{\perp}=I_n-X(X'X)^{-1}X'$, la matrice de projection sur l'espace orthogonal de $E_{p,n}$, où I_n

est la matrice identité $n \times n$.

Posons aussi

$$\mathcal{D}_n^2 = \frac{1}{n} Y_n' \Pi_n^{\perp} Y_n \quad \text{et} \quad \widetilde{\mathcal{D}}_n^2 = \frac{1}{n} f_n' \Pi_n^{\perp} f_n \, .$$

Notons que si $\mathcal{D}_n^2$ est suffisamment petit, on en déduit que f est proche de sa projection sur E_p, ainsi notre statistique de test sera basée sur cette expression empirique.
En remplaçant Y_n par $f_n + \varepsilon_n$, on vérifie que

$$E(\mathcal{D}_n^2) = \widetilde{\mathcal{D}}_n^2 + \frac{n-p}{n} \sigma^2 \, .$$

On est amené ainsi à considérer $\mathcal{D}_n^2 - \frac{n-p}{n}\sigma^2$, mais σ^2 est inconnu. On l'estime à l'aide de l'estimateur introduit par Rice (1984)

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n-1} \left(Y_{i,n} - Y_{i-1,n} \right)^2 \, .$$

On obtient la statistique de test

$$\widehat{\mathcal{D}}_n^2 = \mathcal{D}_n^2 - \frac{n-p}{n} \widehat{\sigma}^2 \, ;$$

on rejette l'hypothèse H_0 : "$f \in E_p$" si

$$\widehat{\mathcal{D}}_n^2 > c_\alpha \, ,$$

où c_α est un nombre réel positif.

Notre résultat principal est le suivant.

Theoreme 1 *Si les conditions* $(A1)$*,* $(A2)$ *et* $(A3)$ *sont satisfaites, alors*

$$\sqrt{n} \left\{ \widehat{\mathcal{D}}_n^2 - \widetilde{\mathcal{D}}_n^2 + \frac{1}{2n} \sum_{i=2}^{n-1} \Big(f(t_{i,n}) - f(t_{i-1,n}) \Big) \right\} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^4 + 4\sigma^2 \mathcal{D}^2\right)$$

où $\mathcal{D}^2 = \mathcal{D}^2(f)$.

Remarque. Sous H_0 , on a $\widetilde{\mathcal{D}}_n^2 = 0$ mais $\dfrac{1}{2\sqrt{n}} \displaystyle\sum_{i=2}^{n-1} \Big(f(t_{i,n}) - f(t_{i-1,n}) \Big)^2$ n'est pas nécessairement nul, ni même négligeable en général.

3 Application

3.1 Test dans un modèle de régression localement höldérienne

On suppose que, dans le modèle de régression étudié, f est une fonction localement höldérienne d'ordre inconnu, (ou seulement Riemann-intégrable) et on considère un espace vectoriel E_p tel que

- (A4) les fonctions $x_1, \ldots, x_p$ sont localement höldériennes d'ordre $\gamma > 1/2$.

On a donc la proposition.

Proposition 1 *Si les conditions* $(A1)$, $(A2)$ *et* $(A4)$ *sont satisfaites on a, sous* H_0,

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

Cette proposition donne le niveau asymptotique du test; le théorème 1 donne la puissance pour des alternatives qui peuvent être höldériennes d'ordre $\delta < 1/2$, (ou seulement Riemann-intégrable).

Bibliographie

[1] Dette, H. and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 778-800.
[2] Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression using nonparametric via order selection criteria. *Ann. Stat.,* **20**, 1412-1425.
[3] 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.
[4] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 1926-1947.
[5] Mohdeb, Z. and Mokkadem, A. (2004). Average squared residuals approach for testing linear hypotheses in nonparametric regression. *J. Nonparametr. Stat.* **16**, no. 1-2, 3–12.
[6] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Stat.* **12**, 1215-1230.