Test de rupture de régression

Zaher Mohdeb

Faculté de Génie des Procédés
Université de Constantine 3, Salah Boubnider
Constantine, Algérie
et
Laboratoire de Mathématiques et Sciences de la Décision
Université frères Mentouri
Constantine, Algérie
E-mail: zaher.mohdeb@univ-constantine3.dz

Résumé. On considère un modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixé; notre objectif est de construire le test de l'hypothèse linéaire contre les alternatives de ruptures de modèle et ce sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative, ce qui inclut la possibilité que les fonctions soient höldériennes d'ordre supérieur à 1/2. On établit la normalité asymptotique de la statistique de test sous l'hypothèse nulle ainsi que sous l'hypothèse alternative de ruptures de modèle.

Mots-clés. Hypothèse linéaire, régression non paramétrique, rupture de modèle.

Abstract. We consider a regression model in the case of a homoscedastic error structure and fixed design, our aim is to build the test of the linear hypothesis versus regime switching models without regularity condition, and also under either the null or the alternative hypotheses, which includes the possibility of functions satisfying the Hölder condition of order greater than 1/2. We establish the asymptotic normality of the test statistic under the null hypothesis and the alternative one.

Keywords. Linear hypothesis, nonparametric regression, regime switching.

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 $t_{i,n}$, $i = 1, \ldots, n$, 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 $g_1, \ldots, g_p$ 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:

$$H_0 : \ f \in E_p \quad \text{contre} \quad H_1 : \begin{cases} \exists\, s \in]0,1[\ \text{ tel que } \ f = \phi \mathbb{1}_{[0,s]} + \psi \mathbb{1}_{]s,1]}, \\ \phi \in E_p\,, \ \ \psi \text{ Riemann intégrable et } f \notin E_p\,. \end{cases} \tag{2}$$

La plupart des travaux sur les tests d'hypothèses dans le modèle (1) supposent des conditions de régularité sur $f, \ \ g_1, \ldots, g_p$; généralement ces fonctions sont supposées höldériennes. On peut citer, sans être exhaustif, Cox et al (1988), Eubank et Spiegelmann (1990), Eubank et Hart (1992), Azzalini et Bowman (1993), Härdle et Mammen (1993). Les tests basés sur l'estimation sur la L^2-distance entre f et E_p sont étudiés par Dette et Munk ((1998), Munk et Dette (1998), Mohdeb et Mokkadem (2004), avec l'hypothèse que f est höldérienne d'ordre $\gamma > 1/2$.

Dans ce travail, on applique l'approche utilisée dans Mohdeb et Mokkadem (2015) et Lessak et Mohdeb (2015) pour construire le test d'hypthèses (2) dans le modèle (1). On suppose que $f, \ g_1, \ldots, g_p$ sont Riemann-intégrables; sous cette seule condition sur les fonctions, on établit la normalité asymptotique de la statistique de test qui permet de construire le test (2) et d'avoir la puissance pour des alternatives de ruptures de modèle.

Dans la section 2, on introduit les hypothèses et on présente notre résultat principal. Des simulations ont été menées pour étudier la performance du test proposée dans la section 3.

2 Hypothèses et résultats

On considère le modèle de régression (1) et E_p est l'espace vectoriel engendré par des fonctions fixées $g_1, \ldots, g_p$ 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, $g_1, \ldots, g_p$ sont Riemann-intégrables.

Remarque. Notons que l'hypothèse (A1) implique que pour toute fonction h Riemann-intégrable

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(t_{i,n}) = \int_0^1 h(t)\, dt.$$

De plus, nous supposons que les fonctions que nous considérons appartiennent également

à $L^2(dt)$, muni de son produit scalaire usuel. On pose,

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

la distance entre f et le sous-espace E_p, $Y := (Y_{1,n}, \ldots, Y_{n,n})'$, $f_n := \Big(f(t_{1,n}), \ldots, f(t_{n,n})\Big)'$, $g_{k,n} := \Big(g_k(t_{1,n}), \ldots, g_k(t_{n,n})\Big)'$, $k = 1, \ldots, p$, et $G := (g_{1,n}, \ldots, g_{p,n})$.

On note aussi $E_{p,n}$, le sous-espace de $\mathbb{R}^n$ engendré par $\{g_{1,n}, \ldots, g_{p,n}\}$ qui est une discrétisation du sous-espace E_p, $\Pi_n = G(G'G)^{-1}G'$, la matrice de projection sur $E_{p,n}$ et $\Pi_n^{\perp} = I_n - G(G'G)^{-1}G'$, la matrice de projection sur l'espace orthogonal de $E_{p,n}$, où I_n est la matrice identité $n \times n$.

On considère la statistique suivante définie par

$$D_n^2 := \frac{1}{n} Y' \Pi_n^{\perp} Y \,.$$

On vérifie que

$$E(D_n^2) = \widetilde{D}_n^2 + \frac{n-p}{n}\sigma^2 \,, \quad \text{où} \quad \widetilde{D}_n^2 = \frac{1}{n} f_n' \Pi_n^{\perp} f_n \,.$$

On est amené ainsi à considérer $D_n^2 - \dfrac{n-p}{n}\sigma^2$, mais σ^2 est inconnu. On l'estime à l'aide de l'estimateur suivant, introduit par Gasser, Sroka, et Jennen-Steinmetz (1986)

$$S_\varepsilon^2 = \frac{1}{6(n-2)} \sum_{i=2}^{n-1} \left(Y_{i+1,n} + Y_{i-1,n} - 2Y_{i,n}\right)^2 \,.$$

On obtient ainsi la statistique de test donnée par

$$\widehat{D}_n^2 = D_n^2 - \frac{n-p}{n} S_\varepsilon^2 \,;$$

et on rejette l'hypothèse H_0 : "$f \in E_p$" si $\widehat{D}_n^2 > u_\alpha$, où u_α est un nombre réel positif.

Notre résultat principal est le suivant.

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

$$\sqrt{n}\left\{\widehat{D}_n^2 - \widetilde{D}_n^2 + B_n(f)\right\} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4 + 4\sigma^2 \mathcal{D}^2(f)\right) \,,$$

où $B_n(f) = \dfrac{1}{6n} \displaystyle\sum_{i=2}^{n-1} \Big(f(t_{i+1,n}) + f(t_{i-1,n}) - 2f(t_{i,n})\Big)^2$.

Remarque. Sous H_0, on a $\widetilde{D}_n^2 = 0$ mais $B_n(f)$ n'est pas nécessairement nul, ni même négligeable en général.

3 Applications

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 $g_1, \ldots, g_p$ sont localement höldériennes d'ordre $\gamma > 1/2$.

Dans ce cas, il existe une subdivision de $[0,1]$ en intervalles $I_1, \ldots, I_q$ et un réel $\delta > 1/2$, (δ est le plus petit des ordres des fonctions $g_1, \ldots, g_p$) tel que tout élément de E_p est Hölder d'ordre δ sur chaque I_j. Cela n'est évidemment pas le cas pour l'alternative, c'est-à-dire l'ensemble des fonctions localement höldériennes qui n'appartiennent pas E_p.

Il est facile de vérifier que pour tout $f \in E_p$,

$$B_n(f) = o\left(\frac{1}{\sqrt{n}}\right).$$

On a donc la proposition.

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

$$\sqrt{n}\widehat{D}_n^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{17}{9}\sigma^4\right).$$

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). En pratique la variance σ^2 des erreurs est généralement inconnue, on peut considérer un estimateur consistent $\widehat{\sigma}^2$ de σ^2. On rejette l'hypothèse nulle H_0 : "$f \in E_p$", si

$$\frac{\sqrt{n}}{\widehat{\sigma}^2}\,\widehat{D}_n^2 \;>\; z_{1-\alpha},$$

où $z_{1-\alpha}$ est le $(1-\alpha)$quantile d'une loi normale standard.

3.2 Test de rupture de modèle

Soit E_p un sous-espace vectoriel engendré par des fonctions $g_1, \ldots, g_p$ localement höldériennes d'ordre $\gamma > 1/2$. On veut tester l'hypothèse:

$$\mathrm{H}_0 : \quad f \in E_p \quad \text{contre} \quad \mathrm{H}_1 : \quad \begin{cases} \exists\, s \in]0,1[\text{ tel que } f = \phi \mathbb{1}_{[0,s]} + \psi \mathbb{1}_{]s,1]} \\ \\ \phi \in E_p\,, \;\; \psi \text{ Riemann-intégrable et } f \notin E_p\,. \end{cases}$$

Comme l'alternative H_1 est contenue dans l'alternative du test de la sous-section 3.1, on peut utiliser ce dernier pour tester la rupture.

4 Simulations

On a réalisé des simulations pour étudier la puissance au niveau $\alpha = 5\%$ du test de l'hypothèse

$$H_0 : \quad f(t) = t \quad \text{contre} \quad H_1 : f(t) = t\mathbb{1}_{[0,s]}(t) + \beta t\mathbb{1}_{]s,1]}(t)\,, \quad \beta \neq 1\,.$$

On a considéré un échantillonnage régulier, $t_{i,n} = \frac{i-1}{n-1},\ i = 1, \ldots, n;$ avec $n = 64$.

L'hypothèse H_0 est rejetée si

$$\left(\frac{9n}{17}\right)^{1/2} \frac{\widehat{D}_n^2}{\widehat{\sigma}^2} > 1.65\,,$$

avec

$$\widehat{D}_n^2 = \frac{1}{n}\sum_{i=1}^{n} |Y_{i,n} - \widehat{a}t_{i,n}|^2 - \frac{n-1}{n}S_\varepsilon^2$$

et

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_{i,n} - \widehat{a}t_{i,n})^2,$$

où

$$\widehat{a} = \frac{\sum_{i=1}^{n} t_{i,n}Y_{i,n}}{\sum_{i=1}^{n} t_{i,n}^2}\,.$$

Pour s et β, on a considéré les valeurs $s = 0.238$, 0.492, 0.619, 1.000 et $\beta = 0.00$, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00.

On a aussi considéré plusieurs valeurs de la variance σ^2 du bruit: $\sigma = 0.05$, 0.10, 0.20 et 0.50.

L'analyse des résultats obtenus montre que pour des petites variances de ε, l'hypothèse H_0 est rejetée avec une proportion proche de 1.

Bibliographie

[1] Azzalini, A. and Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. *J. Roy. Statist. Soc. Ser. B*, **55**, 549-557.
[2] Cox, D., Koh, G., Wahba, G. and Yandell, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, **16** 113-119.

[3] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 2, 778-800.
[4] Eubank, R. L. and Hart, J. D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Stat.*, **20**, 3, 1412-1425.
[5] 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.
[6] Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika,* **73**, 625-633.
[7] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 4, 1926-1947.
[8] Lessak, R. and Mohdeb, Z. (2015). Testing the linear regression model null hypothesis versus regime switching alternatives. *Afr. Stat.*, **10**, 807-813.
[9] Mohdeb, Z. and Mokkadem, A. (2004). Average squared residuals approach for testing linear hypothesis in nonparametric regression. *J. Nonparametric Stat.*, **16**, 1-2, 3-12.
[10] Mohdeb, Z. and Mokkadem, A. (2015). Testing linear regression models in non regular case. *Comm. Statist. Theory Methods*, **44**, 21, 4476-4490.
[11] Munk, A. and Dette, H. (1998). Nonparametric comparison of several regression functions: exact and asymptotic theory. *Ann. Stat.*, **26**, 6, 2339-2368.