Un test de détection de rupture dans un modèle de régression

Zaher Mohdeb

Ecole Nationale Polytechnique de Constantine, Algérie
et
Laboratoire de Mathématiques et Sciences de la Décision, Université frères Mentouri, Constantine, Algérie
E-mail: zaher.mohdeb@umc.edu.dz, z.mohdeb@gmail.com

Résumé. On considère un modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixé, notre but est de construire le test de l'hypothèse d'un modèle de régression 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. 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 hypothesis test of the linear regression model versus regime switching models without regularity condition, and also under either the null or the alternative hypotheses. 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

$$Z_{i,n} = g(t_{i,n}) + \varepsilon_{i,n}\ , \quad i = 1, \dots, n\,, \tag{1}$$

où g est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{i,n}$, $i = 1, \dots, n$, 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 $x_1, \dots, x_p$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit U_p l'espace vectoriel engendré par $x_1, \dots, x_p$. On veut tester l'hypothèse:

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

La plupart des travaux sur les tests d'hypothèses dans le modèle (1) supposent des conditions de régularité sur g, $x_1, \ldots, x_p$; 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 U_p sont étudiés par Dette et Munk (1998), Munk et Dette (1998), Mohdeb et Mokkadem (2004), avec l'hypothèse que g 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 g, $x_1, \ldots, x_p$ 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.

2 Hypothèses et résultats

On considère le modèle de régression (1) et U_p est l'espace vectoriel engendré par des fonctions fixées $x_1, \ldots, x_p$ définies sur $[0, 1]$ et linéairement indépendantes.

Nos hypothèses sont les suivantes:

- (A1) $\max_{i=2,\ldots,n} \left| (t_{i,n} - t_{i-1,n}) - \frac{1}{n} \right| = o\left(\frac{1}{n}\right)$;
- (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) La fonction g est Riemann-intégrable.
- (A4) Les fonctions $x_1, \ldots, x_p$ sont localement höldériennes d'ordre $\gamma > 1/2$.

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

$$\mathcal{M}^2(g) := \min_{v \in U_p} \|g - v\|^2 \tag{3}$$

la distance entre g et le sous-espace U_p, $Z := (Z_{1,n}, \ldots, Z_{n,n})'$, $g_n := \Big(g(t_{1,n}), \ldots, g(t_{n,n})\Big)'$, $x_{k,n} := \Big(x_k(t_{1,n}), \ldots, x_k(t_{n,n})\Big)'$, $k = 1, \ldots, p$, et $X := (x_{1,n}, \ldots, x_{p,n})$.
On note aussi $U_{p,n}$, le sous-espace de $\mathbb{R}^n$ engendré par $\{x_{1,n}, \ldots, x_{p,n}\}$ qui est une discrétisation du sous-espace U_p, $\Pi_n = X(X'X)^{-1}X'$, la matrice de projection sur $U_{p,n}$ et $\Pi_n^\perp = I_n - X(X'x)^{-1}X'$, la matrice de projection sur l'espace orthogonal de $U_{p,n}$, où I_n est la matrice identité $n \times n$.

On considère la statistique suivante définie par

$$\mathcal{M}_n^2 := \frac{1}{n} Z' \Pi_n^{\perp} Z \,. \tag{4}$$

On vérifie que

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

On estime σ^2 à l'aide de l'estimateur suivant, introduit par Rice (1984)

$$\widehat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} \left(Z_{i,n} - Z_{i-1,n}\right)^2 \,. \tag{5}$$

On obtient ainsi la statistique de test donnée par

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

et on rejette l'hypothèse H_0 : $"g \in U_p"$ si $\widehat{\mathcal{M}}_n^2 > u_\alpha$, 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*

$$\sqrt{n}\left\{\widehat{\mathcal{M}}_n^2 - \widetilde{\mathcal{M}}_n^2 + D_n(g)\right\} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^4 + 4\sigma^2\mathcal{M}^2(g)\right)\,,$$

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

Nous avons le résultat suivant, comme conséquence du Théorème 1, sous l'hypothèse nulle H_0.

Corollaire 1 *Si les conditions* (A1)-(A4) *sont satisfaites, alors sous l'hypothèse nulle* H_0*, on a*

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

Ce corollaire donne le niveau asymptotique du test et le Théorème 1 donne la puissance du test pour les hypothèses alternatives de ruptures de modèles. 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 : $"g \in U_p"$, si

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

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

3 Simulations

Dans nos simulations, on étudie la puissance au niveau $\alpha = 5\%$ du test de l'hypothèse

$$\mathrm{H}_0 : \quad g(t) = t \quad \text{contre} \quad \mathrm{H}_1 : g(t) = t\mathbb{1}_{[0,s]}(t) + \beta t \mathbb{1}_{]s,1]}(t)\,, \quad \beta \neq 1\,.$$

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.$\mathcal{N}(0, \sigma^2)$, $n = 64$ et pour différentes valeurs respectivement de s, β et σ^2.
L'hypothèse H_0 est rejetée si $\dfrac{\sqrt{n}}{\widehat{\sigma}^2}\,\widehat{\mathcal{M}}_n^2 > 1.65$, avec $\widehat{\mathcal{M}}_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} |Z_{i,n} - \widehat{a}t_{i,n}|^2 - \dfrac{n-1}{n}\widehat{\sigma}_R^2$,
où $\widehat{a} = \dfrac{\sum_{i=1}^n t_{i,n} Z_{i,n}}{\sum_{i=1}^n t_{i,n}^2}$.
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.