Non parametric Tests for Detecting Random Criterion in Regression Model

Mohamed Fihri[1], Abdelhadi Akharif[1] & Amal Mellouk[2]

[1] *Laboratoire de Mathématiques et Applications, Faculté des Sciences et Techniques de Tanger - Maroc.* `m.fihri@gmail.com` & `aakharif@gmail.com`
[2] *Centre Régional des Métiers de l'Education et de la Formation de Tanger - Maroc.* `melloukamal@hotmail.com`

Résumé. Dans cet article, on s'intéresse au problème de tester l'hypothèse nulle d'un modèle de régression linéaire standard (SLR) contre l'alternative d'un modèle de régression à coefficient aléatoire (RCR). Pour cela, on établit la normalité locale asymptotique (LAN), qui nous permettra de construire des tests localement et asymptotiquement optimaux. On donne la version basée sur les rangs et les rangs signés. Ces tests sont les plus puissants et valable pour une large classe de densités. Ce qui n'est pas le cas pour le test de Wilcoxon proposé par Ramanathan and Rajarshi (1992). Puisqu'il est valide uniquement sous des densités symétriques. Les simulations confirment la performance des tests proposés.

Mots-clés. Propriété LAN, test optimal, *RCR model*, tests de rangs signés et non signés.

Abstract. We consider the problem of detecting the randomness criterion in a regression model. This is the problem of testing a Standard Linear Regression (*SLR model*) against an alternative with Random Coefficient Regression model (*RCR model*). We propose a non parametric (rank and signed rank) test locally and asymptotically optimal in the Hajek sense. Simulations confirm the performance of the given tests.

Keywords. LAN Property, test optimal, *RCR model*, rank test, signed rank test.

1 Introduction

On s'intéresse au modèle de régression à coefficient aléatoire (RCR) définie par

$$Y_i = \mu + (\beta + \xi_i)X_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

avec

▷ Y_i désigne la réponse observée pour l'individu i, X_i est la variable explicative exogène déterministe (l'observation) et μ et β sont les paramètres de régression,

- ▷ $(\varepsilon_i)_{1\leq i\leq n}$ est une suite de variables aléatoires indépendantes et identiquement distribuées *i.i.d.* avec $\mathbb{E}(\varepsilon_i) = 0$, $0 < \mathbb{E}(\varepsilon_i^2) = \sigma^2 < \infty$ et de densité $\varepsilon \mapsto f(\varepsilon) := (1/\sigma)f_1(\varepsilon/\sigma)$,
- ▷ $(\xi_i)_{1\leq i\leq n}$ est une suite de variables aléatoires *i.i.d.* $(0, \sigma_\xi^2)$ de densité $\xi \mapsto h(\xi) := (1/\sigma_\xi)h_1(\xi/\sigma_\xi)$, et indépendante de $(\varepsilon_i)_{1\leq i\leq n}$.

Notons que si $\sigma_\xi^2 = 0$, alors le modèle RCR se réduit au modèle linéaire SLR :

$$Y_i = \mu + \beta X_i + \varepsilon_i, \qquad i = 1, \dots, n. \tag{2}$$

Parmi les travaux qui se sont intéressés à cette problématique, on cite : Newbold et Bos (1985) qui donnent le test du multiplicateur de Lagrange, Ramanathan et Rajarshi (1992) qui proposent le test de rang signé basé sur les scores de Wilcoxon. En se basant sur les travaux de Akharif et Hallin (2003) et Bennala et *al.* (2012), nous proposons une procédure de test non paramétrique, basée sur les rangs, localement et asymptotiquement optimale au sens de Hájek. On donne la version basée sur les rangs (signés et non signés) de *van der Waerden* (score normal) et de *Wilcoxon* (score logistique). A partir des simulations, on montre, pour des densités asymétriques (skew-normal et skew-student), la supériorité des tests de rangs par rapport au test proposé par Ramanathan et Rajarshi.

2 Plan de la présentation

Pour tester la présence d'un *critère aléatoire* dans un modèle de régression, il revient de tester :

$$\begin{cases} \mathcal{H}_0 : & \sigma_\xi^2 = 0 \\ \mathcal{H}_1 : & \sigma_\xi^2 > 0 \end{cases} \Longleftrightarrow \begin{cases} \mathcal{H}_0 : SLR \\ \mathcal{H}_1 : RCR \end{cases}.$$

Le résultat clé pour construire des tests optimaux est de montrer que le modèle étudié jouit de la propriété LAN *(la normalité locale asymptotique)*. On établie la propriété LAN pour la famille des distributions

$$\mathcal{P}_{f_1,h_1}^{(n)} := \left\{ \mathbb{P}_{\sigma_\xi^2;f_1,h_1}^{(n)} : \sigma_\xi^2 \geq 0 \right\}$$

en $\sigma_\xi^2 = 0$. On considère une suite locale d'alternative de la forme $\left(0 + n^{-1/2}K^{(n)}\tau\right)$, où $K^{(n)} = \left(\sum_{i=1}^n X_i^4\right)^{-1/2}$ et $\tau \in \mathbb{R}^+$.

Le problème de test devient : $\begin{cases} \mathcal{H}_0 : \mathbb{P}_{0;f_1}^{(n)} =: \mathbb{P}_{f_1}^{(n)} & : \quad \tau = 0 \\ \mathcal{H}_1 : \mathbb{P}_{\sigma_\xi^2;f_1,h_1}^{(n)} & : \quad \tau > 0 \end{cases}.$

Definitions et Notations 2.1. ▷ *Considérons $\mathcal{F}_0$ (resp. $\mathcal{F}_0^+$) l'ensemble des densités standardisées (resp. l'ensemble des densités standardisées symétriques) définie par :*

$$\mathcal{F}_0 := \left\{ f_1 : \int_{-1}^{1} f_1(z)dz = 0.5 = \int_{-\infty}^{0} f_1(z)dz \right\} \qquad et$$

$$\mathcal{F}_0^+ := \left\{ f_1 : f_1(-z) = f_1(z) \ and \ \int_{-1}^{1} f_1(z)dz = 0.5 = \int_{-\infty}^{0} f_1(z)dz \right\}.$$

Notons que, sous $\mathcal{F}_0$ et $\mathcal{F}_0^+$ la médiane et l'écart absolu médian sont respectivement 0 et σ. Cette normalisation qui, contrairement à l'habituel basée sur la moyenne et l'écart-type, permet d'éviter toutes les hypothèses sur les moments et n'a aucun impact sur les résultats ultérieurs ;

- ▷ *Supposons que $f_1 > 0$, de classe $\mathcal{C}^2$ et de derivée seconde $\ddot{f}_1$;*
- ▷ *Posons $\psi_{f_1} := \ddot{f}_1 / f_1$ et supposons que $\mathcal{I}_\psi(f_1) := \int_{\mathbb{R}} \psi_{f_1}^2(z) f_1(z)dz < \infty$;*
- ▷ *On définit les résidus standardisés*

$$Z_i := \sigma^{-1}(Y_i - \mu - \beta X_i), \qquad i = 1, ..., n.$$

On note que, sous l'hypothèse nulle $\mathbb{P}_{f_1}^{(n)}$, ces résidus coïncident avec $\dfrac{\varepsilon_i}{\sigma}$.

On obtient donc la proposition suivante :

Proposition 2.1 (LAN). *(i) La famille $\mathbb{P}_{f_1,h_1}^{(n)}$ est LAN en 0, avec comme suite centrale*

$$\Delta_{f_1}^{(n)} := \frac{1}{2\sigma^2\sqrt{n}} \sum_{i=1}^{n} \psi_{f_1}(Z_i) K^{(n)} X_i^2 \tag{3}$$

et comme variance

$$\gamma_{f_1} := \frac{1}{4n\sigma^4} \mathcal{I}_\psi(f_1), \tag{4}$$

(ii) pour tout $\tau \in \mathbb{R}^+$, on a, sous $\mathbb{P}_{f_1}^{(n)}$,

$$\begin{aligned} \Lambda_{n^{-1/2}K^{(n)}\tau/0;f_1,h_1}^{(n)} :=& \log \left(\frac{d\mathbb{P}_{n^{-1/2}K^{(n)}\tau;f_1,h_1}^{(n)}}{d\mathbb{P}_{f_1}^{(n)}} \right) \\ =& \tau \Delta_{f_1}^{(n)} - \frac{1}{2}\tau^2 \gamma_{f_1} + o_{\mathbb{P}}(1), \end{aligned}$$

(iii) $\Delta_{f_1}^{(n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \gamma_{f_1})$, sous $\mathbb{P}_{f_1}^{(n)}$.

Ce résultat nous amène à construire des tests *paramétriques* localement et asymptotiquement optimal sous une densité f_1 spécifiée et des tests *non paramétriques* basés sur les *rangs signés et non signés.*

Remark 2.1. *La densité h_1 sert seulement à la définition de la vraisemblance. Elle n'a pas d'influence sur l'établissement du test.*

La conséquence principale de la propriété LAN est la convergence de la suite d'expériences locales considérée vers le problème de *position Gaussien classique de forme* $\mathcal{N}(\gamma\tau, \gamma)$. Le test est basé sur $\Delta_{f_1}^{(n)}$, donc sur $T_{f_1}^{(n)}$, avec

$$\begin{aligned} T_{f_1}^{(n)} &:= (\gamma_{f_1})^{-1/2} \Delta_{f_1}^{(n)} \\ &= \frac{K^{(n)}}{\sqrt{\mathcal{I}_\psi(f_1)}} \sum_{i=1}^{n} \psi_{f_1}(Z_i) X_i^2. \end{aligned} \tag{5}$$

On déduit donc le résultat suivant :

Proposition 2.2. *(i) $T_{f_1}^{(n)}$ est asymptotiquement normale, de moyenne zéro sous $\mathbb{P}_{f_1}^{(n)}$, de moyenne $\gamma_{f_1}^{1/2}\tau$ sous $\mathbb{P}_{n^{-1/2}K^{(n)}\tau;f_1,h_1}^{(n)}$ et de variance 1 sous les deux hypothèses ;*

(ii) le test rejetant l'hypothèse nulle $\mathcal{H}_0^{(n)}$ quand

$$T_{f_1}^{(n)} > z_{1-\alpha},$$

est localement asymptotiquement le plus puissant au niveau α, contre l'alternative d'une régression à coefficient aléatoire.

Ci dessous, on donne les versions basées sur les rangs de la statistique du test

- ***Statistique de van der Waerden (score normal)***

$$T_{vdW}^{(n)} = \left[\frac{1 - \frac{1}{n}}{s_{vdW}^{2(n)} \sum_{i=1}^{n} \left(X_i^2 - \overline{X^2} \right)^2} \right]^{1/2} \sum_{i=1}^{n} X_i^2 \left\{ \left(\Phi^{-1}\left(\frac{R_i^{(n)}}{n+1} \right) \right)^2 - \overline{\psi}_{vdW}^{(n)} \right\}, \tag{6}$$

avec $\overline{\psi}_{vdW}^{(n)} := \dfrac{1}{n} \sum_{i=1}^{n} \left(\Phi^{-1}\left(\dfrac{i}{n+1} \right) \right)^2$ et

$$\begin{aligned} s_{vdW}^{2(n)} &:= \frac{n-1}{n^2} \sum_{i=1}^{n} \left[\left(\Phi^{-1}\left(\frac{i}{n+1} \right) \right)^2 - 1 \right]^2 \\ &- \frac{1}{n^2} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i\neq j}}^{n} \left[\left(\Phi^{-1}\left(\frac{i}{n+1} \right) \right)^2 - 1 \right] \left[\left(\Phi^{-1}\left(\frac{j}{n+1} \right) \right)^2 - 1 \right]. \end{aligned}$$

- ***Statistique de Wilcoxon (score logistique)***

$$T_W^{(n)} = \left[\frac{1-\frac{1}{n}}{s_W^{2(n)} \sum_{i=1}^n \left(X_i^2 - \overline{X^2} \right)^2} \right]^{1/2} \sum_{i=1}^n X_i^2 \left\{ 6 \left(\frac{R_i^{(n)}}{n+1} \right)^2 - 6 \left(\frac{R_i^{(n)}}{n+1} \right) - \overline{\psi}_W^{(n)} \right\}, \quad (7)$$

avec $\overline{\psi}_W^{(n)} = -\frac{n+2}{n+1}$ et $s_W^{2(n)} = \frac{(n-2)(n-1)(n+2)}{5(n+1)^3}$.

Extrait des Simulations

Pour évaluer la performance des tests proposés, considérons le modèle suivant :

$$Y_i = \mu + \beta X_i + \xi_i X_i + \varepsilon_i, \qquad i = 1, \ldots, n = 100, \text{ avec}, \quad (8)$$

(a) $\mu = 1$ et $\beta = 10$;
(b) les X_i sont *i.i.d.* uniforme sur $(0, 10)$;
(c) les ξ_i sont *i.i.d.* Gaussiens de moyenne nulle et d'écart-type $\sigma_\xi = 0$ (**hypothèse nulle**), 0.05, 0.1, 0.15, 0.2 and 0.25 (**hypothèses alternatives**) ;
(d) les ε_i sont *i.i.d.* de densité asymétrique – skew-normal ($s\mathcal{N}$), skew-Student t_5 (st_5).

On génère $N = 2500$ réplications de taille $n = 100$ de (8). Les résultats sont présentés dans la table 1.

g_1	$Test$	σ_ξ					
		0	0.05	0.1	0.15	0.2	0.25
$s\mathcal{N}$	T_{vdW}	0.0468	0.2032	0.5804	0.8720	0.9624	0.9900
	T_W	0.0436	0.1392	0.4808	0.8132	0.9348	0.9804
	T_{t_1}	0.0444	0.0400	0.0692	0.1476	0.2248	0.3148
	T_{t_3}	0.0508	0.1336	0.4348	0.7680	0.9228	0.9748
	T_{t_5}	0.0420	0.1208	0.4420	0.7876	0.9168	0.9768
	T_{RR}	0.0500	0.0728	0.1600	0.3776	0.6116	0.8108
st_5	T_{vdW}	0.0484	0.2088	0.6092	0.8660	0.9596	0.9924
	T_W	0.0500	0.1392	0.4900	0.8056	0.9380	0.9848
	T_{t_1}	0.0520	0.0392	0.0800	0.1388	0.2380	0.3224
	T_{t_3}	0.0500	0.3980	0.3856	0.6792	0.8596	0.9432
	T_{t_5}	0.0484	0.1192	0.4508	0.7700	0.9216	0.9772
	T_{RR}	0.0552	0.0760	0.1632	0.3740	0.6132	0.8072

Table 1 – Rejection frequencies (out of 2500 replications) for $\sigma_\xi = 0$ (null hypothesis), 0.05, 0.1, 0.15, 0.2, 0.25 (alternative hypotheses), with error density g_1 that is skew-normal density ($f_{s\mathcal{N}}$) and skew-Student density with 5 $d.f$ (st_5), of the van der Waerden test (T_{vdW}), the Wilcoxon test (T_W), the Student T_{t_ν} tests (t_ν-score with $\nu = 1, 3, 5$) and the Ramanathan and Rajarshi test (T_{RR}), for $n = 100$, at asymptotic level $\alpha = 5\%$.

Bibliographie

[1] Akharif, A., Hallin, M. (2003) *Efficient detection of random coefficients in autoregressive models*, The Annals of Statistics **31**, p. 675-704.

[2] Bennala, N., Hallin, M., Paindaveine, D. (2012) *Pseudo-Gaussian and rank-based optimal tests for random individual effects in large n small T panels*, Journal of Econometrics **170**, p. 50-67.

[3] Hájek, J., Šidák, Z., Sen, P.K., (1999), *Theory of Rank Tests*, Academic Press, New York-London 2nd edition

[4] Le Cam, L.M. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

[5] Newbold, P., Bos, T. (1985) *Stochastic parameter regression models*, Series : Quantitative applications in the Social Sciences.

[6] Ramanathan, T.V., Rajarshi, M.B. (1992) *Rank Tests for Testing Randomness of a Regression Coefficient in a Linear Regression Model*, Metrika **39**, p. 113-124.

[7] Swensen, A.R. (1985) *The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend*, Journal of Multivariate Analysis **16**, p. 54-70.

[8] van der Vaart, A.W. (1998) *Asymptotic Statistic*, Cambridge University Press, Cambridge.