Afrika Statistika
Vol. 10(1), 2015, pages 807–813.
DOI: http://dx.doi.org/1016929.as.2015.807.71

Afrika Statistika

ISSN 2316-090X

Testing the linear regression model null hypothesis versus regime switching alternatives

Radia Lessak and Zaher Mohdeb[1]

Université frères Mentouri Constantine. Laboratoire de Mathématiques et Sciences de la Décision. Constantine, Algérie

Received January 31, 2015; Accepted October 18, 2015

Abstract. In this paper, we develop a test for the nonparametric regression model in the case of a homoscedastic error structure and fixed design. More precisely, a new statistic is proposed for testing linear hypothesis versus regime switching alternatives; 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. A simulation study is conducted to investigate the finite sample properties of the proposed test.

Résumé. L'objet du présent travail est de construire des tests d'hypothèses dans le modèle de régression non paramétrique à erreurs homoscédastiques et un échantillonnage fixé. Plus précisément, une nouvelle statistique de test est proposée pour construire le test de l'hypothèse linéaire contre un modèle de rupture; sans condition de régularité sur la fonction de régression aussi bien sous l'hypothèse nulle que sous l'alternative. Nous établissons la normalité asymptotique de la statistique de test sous l'hypothèse nulle ainsi que sous l'hypothèse alternative. Une étude de simulation est menée pour étudier les propriétés du test dans le cas de petite taille d'échantillon.

Key words: Linear hypothesis; Nonlinear regression; Nonparametric regression; Regime switching.
AMS 2010 Mathematics Subject Classification : 62G08; 62G10; 62G20.

1. Introduction

We consider the following homoscedastic regression model

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

[1]Corresponding author Zaher Mohdeb: zaher.mohdeb@umc.edu.dz
Radia Lessak : lessak radia@yahoo.fr

where g is an unknown real function, defined on the interval $[0,1]$ and $t_{i,n}$, $i = 1,\dots,n$, is a fixed sampling of the interval $[0,1]$. The errors $\varepsilon_{i,n}$ form a triangular array of random variables with expectation zero and finite variance σ^2.
Let $U_p = \text{span}\{f_1,\dots,f_p\}$, where $f_1,\dots,f_p$ denote p linearly independent functions defined on $[0,1]$. We consider the problem of testing the hypothesis

$$H_0 :\ g \in U_p \quad \text{versus} \quad H_1 :\ \begin{cases} \exists\, s \in]0,1[\ \text{ such that } \ g = \phi \mathbb{I}_{[0,s]} + \psi \mathbb{I}_{]s,1]}, \\ \phi \in U_p\,, \ \ \psi \text{ Riemann integrable and } g \notin U_p\,. \end{cases} \tag{2}$$

Most of the literature for the problem of testing hypotheses in the model (1) assume regularity conditions on $f_1,\dots,f_p$ and the regression function g and generally these functions are assumed to satisfy the Hölder condition. In this model, various tests for the hypotheses were suggested, mostly motivated by assessing the goodness of fit of a specific linear model. We refer to the work of Cox *et al.* (1988), Eubank and Spiegelman (1990), Eubank and Hart (1992), Azzalini and Bowman (1993), Härdle and Mammen (1993). The tests based on the estimation of the L^2-distance between f and U_p have been studied by Dette and Munk (1998), Munk and Dette (1998), Mohdeb and Mokkadem (2004), with the assumption that g satisfies the Hölder condition of order $\gamma > 1/2$. You and Chen (2005), Dette and Marchlewski (2008) and Dette and Hetzler (2009) proposed a test for homoscedasticity in the model (1).

The aim of this paper is to provide a test for the problem (2) in the model (1) without regularity condition on the regression function, and also under either the null or the alternative hypotheses. We only assume that g, $f_1,\dots,f_p$ are Riemann integrable; with this condition on these functions, we establish a theorem of convergence in distribution of the test statistic that allows to construct hypotheses test (2).

The paper is organized as follows. In Section 2 we define the test statistic and present our main result. In Section 3 we report a small simulation study on empirical power of our test. Finally proofs of theoretical results are given in Section 4.

2. Assumptions and results

Consider the regression model (1) and $U_p = \text{span}\{f_1,\dots,f_p\}$, where $f_1,\dots,f_p$ denote p linearly independent functions defined on the interval $[0,1]$.

Our assumptions are the following

- (A1) $\max\limits_{i=2,\dots,n} \left|(t_{i,n} - t_{i-1,n}) - \dfrac{1}{n}\right| = o\left(\dfrac{1}{n}\right)$;
- (A2) $\forall n$, $\varepsilon_{1,n},\dots,\varepsilon_{n,n}$ are independent and $\exists C \in \mathbb{R}^+$ such that $E(\varepsilon_{i,n}^4) < C$, $\forall i$;
- (A3) The function g is Riemann integrable;
- (A4) The functions $f_1,\dots,f_p$ are locally Hölder continuous of order $\gamma > 1/2$.

Functions that we consider, are in $L^2(dt)$ equipped with the usual inner product. We use the following notations

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

is the distance between g and the subspace U_p,

$$Z := (Z_{1,n},\dots,Z_{n,n})'\,,\ \varepsilon := (\varepsilon_{1,n},\dots,\varepsilon_{n,n})'\,,\ g_n := \Big(g(t_{1,n}),\dots,g(t_{n,n})\Big)'\,,$$

$$f_{k,n} := \Big(f_k(t_{1,n}), \ldots, f_k(t_{n,n})\Big)', \quad k = 1, \ldots, p\,, F := (f_{1,n}, \ldots, f_{p,n}),$$

and let $U_{p,n}$ be the subspace of $\mathbb{R}^n$ generated by $\{f_{1,n}, \ldots, f_{p,n}\}$ which is the discretization of the subspace U_p, $\Pi_n = F(F'F)^{-1}F$ and $\Pi_n^{\perp} = I_n - F(F'F)^{-1}F'$, where I_n is the identity matrix.

Now, we consider the following statistic defined by

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

and we check that

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

This suggests to estimate $\mathcal{M}^2(g)$ by $\mathcal{M}_n^2 - \dfrac{n-p}{n}\,\sigma^2$.

Since σ^2 is unknown, we replace it by its consistent estimator $\widehat{\sigma}_R^2$ given by Rice (1984), and defined by

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

The test statistic is then

$$\widehat{\mathcal{M}}_n^2 := \mathcal{M}_n^2 - \frac{n-p}{n}\,\widehat{\sigma}_R^2 \tag{6}$$

and we reject the null hypothesis $H_0:\ g \in U_p$ if $\widehat{\mathcal{M}}_n^2$ exceeds a critical value.

We have the following asymptotic result.

Theorem 1. *Let the assumptions* $(A1)$-$(A3)$ *be fulfilled. Then*

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

where

$$D_n(g) := \frac{1}{2n} \sum_{i=2}^{n-1} \Big(g(t_{i,n}) - g(t_{i-1,n})\Big)^2\,.$$

Remark. Note that under the null hypothesis $H_0 : g \in U_p = \text{span}\{f_1, \ldots, f_p\}$, with $f_1, \ldots, f_p$ locally Hölder continuous of order $\gamma > 1/2$, there exists a subdivision on $[0,1]$ into intervals $I_1, \ldots, I_s$ and a real number $\delta > 1/2$ such that any element of U_p is Hölder continuous of order δ on each I_j, $j = 1, \ldots, s$. In this framework, under H_0, we have $\mathcal{M}^2(g) = 0$, $\widetilde{\mathcal{M}}_n^2 = 0$ and it is easy to check that for all $g \in U_p$, we have $D_n(g) = o\left(\frac{1}{\sqrt{n}}\right)$. Then we have the following corollary as consequence of Theorem 1, under the null hypothesis H_0.

Corollary 1. *Let the assumptions* $(A1)$-$(A4)$ *be fulfilled. Then under the null hypothesis* H_0*, we have*

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

This corollary gives the asymptotic level of the test and Theorem 1 gives the power for the alternatives. The variance error σ^2 is generally unknown in practice, then in order to obtain a test for H_0, we need a consistent estimator $\widehat{\sigma}^2$ of σ^2. We reject the null hypothesis $H_0 : g \in U_p$ if

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

where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution.

3. Simulations

In this section we study the small sample properties of the power of the proposed test. Some simulations are conducted to test the null hypothesis

$$H_0 \;:\; g(t) = at\mathbb{1}_{[0,1]}(t) \;\text{ against }\; H_1 \;: g(t) = at\mathbb{1}_{[0,s]}(t) + \delta t\mathbb{1}_{]s,1]}(t)\,,$$

at the significance level $\alpha = 0.05$, with a nonzero real.

The test statistic is defined by

$$\widehat{\mathcal{M}}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Z_{i,n} - \widehat{\beta}t_{i,n}\right)^2 - \frac{n-1}{n}\,\widehat{\sigma}_R^2\,, \quad \text{where} \quad \widehat{\beta} = \frac{\sum_{i=1}^{n} t_{i,n}Z_{i,n}}{\sum_{i=1}^{n} t_{i,n}^2}\,.$$

The null hypothesis H_0 is rejected if

$$\frac{\sqrt{n}}{\widehat{\sigma}_R^2}\,\widehat{\mathcal{M}}_n^2 \;>\; z_{0.95}\,,$$

where $z_{0.95} = 1.65$ is the 0.95th quantile of the standard normal distribution and $\widehat{\sigma}_R^2$ is the estimator defined in (5).

We consider a sample of size n from model $Z_{i,n} = g(t_{i,n}) + \varepsilon_{i,n}$ for $i = 1, \ldots, n$, where the errors $\varepsilon_{i,n}$ are i.i.d.$\mathcal{N}(0, \sigma^2)$. We use the regular sequence $t_{i,n} = \frac{i-1}{n-1}$, $i = 1, \ldots, n$, on the interval $[0, 1]$.

For s and δ, we consider the values $s = 0.25,\ 0.50, 0.75, 0.90, 0.95, 1.00$ and δ from 0 to 1 with step 0.10. We have also considered the following values of the standard deviation error $\sigma = 0.05$ and $\sigma = 0.10$. For each value of s, δ and σ^2, we replicate the experiment 1000 times and record the proportion of rejection. The results of the empirical power study are shown in Table 1, with $a = 0.75$ and sample size $n = 64$.

Examination of the results given in Table 1 reveals that the empirical power is close to 1 for very small standard deviation ($\sigma \leq 0.05$) and $\delta < 0.40$ and also $s = 0.5, 0.75, 0.90, 0.95$. We note also that for $\delta = 0.70$ or $s = 1$, the empirical power is close to significance level $\alpha = 0.05$ since in this case, the alternative is also close to the null hypothesis.

σ	δ	$s = 0.25$	$s = 0.50$	$s = 0.75$	$s = 0.90$	$s = 0.95$	$s = 1.00$
0.05	0.00	0.993	1.000	1.000	1.000	1.000	0.054
	0.10	0.977	1.000	1.000	1.000	1.000	0.062
	0.20	0.914	1.000	1.000	1.000	1.000	0.064
	0.30	0.719	1.000	1.000	1.000	1.000	0.044
	0.40	0.516	1.000	1.000	1.000	1.000	0.069
	0.50	0.244	0.990	1.000	1.000	0.994	0.053
	0.60	0.118	0.623	0.976	0.881	0.685	0.057
	0.70	0.064	0.113	0.174	0.091	0.076	0.058
	0.80	0.071	0.092	0.141	0.095	0.070	0.046
	0.90	0.080	0.614	0.963	0.869	0.597	0.050
	1.00	0.170	0.983	1.000	1.000	0.992	0.056
0.10	0.00	0.509	1.000	1.000	1.000	1.000	0.053
	0.10	0.358	1.000	1.000	1.000	1.000	0.050
	0.20	0.244	0.999	1.000	1.000	1.000	0.040
	0.30	0.172	0.953	1.000	1.000	0.974	0.037
	0.40	0.111	0.804	0.993	0.979	0.861	0.041
	0.50	0.082	0.423	0.834	0.747	0.494	0.039
	0.60	0.048	0.151	0.326	0.276	0.177	0.038
	0.70	0.038	0.055	0.066	0.062	0.064	0.043
	0.80	0.049	0.048	0.066	0.059	0.057	0.055
	0.90	0.055	0.148	0.321	0.257	0.163	0.053
	1.00	0.061	0.423	0.833	0.728	0.450	0.041

Table 1. Proportion of rejection in 1000 samples of size $n = 64$.

4. Proofs

Proof of Theorem 1. First, let us define

$$\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)' := (F'F)^{-1}F'Z\,, \quad \widehat{\widehat{\beta}} = (\widehat{\widehat{\beta}}_1, \ldots, \widehat{\widehat{\beta}}_p)' := (F'F)^{-1}F'\varepsilon\,,$$
$$\widetilde{\beta} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)' := (F'F)^{-1}F'g_n\,,$$

where F, Z, ε and g_n are defined in Section 2.

That means that the projections of Z, ε and g_n, on $U_{p,n}$, are given by $\Pi_n Z = F\widehat{\beta}$, $\Pi_n\varepsilon = F\widehat{\widehat{\beta}}$ and $\Pi_n g_n = F\widetilde{\beta}$, respectively, where $U_{p,n}$ and Π_n are defined in Section 2.

We check easily that

$$\widetilde{\mathcal{M}}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left|g(t_{i,n}) - \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right|^2.$$

Now, a straightforward calculation shows

$$\widehat{\mathcal{M}}_n^2 - \widetilde{\mathcal{M}}_n^2 = V_n^2 - \frac{n-p}{n}\widehat{\sigma}_R^2 + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_{i,n}\left(g(t_{i,n}) - \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right),$$

where

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left|Z_{i,n} - \sum_{k=1}^{p}\widehat{\beta}_k f_k(t_{i,n}) - g(t_{i,n}) + \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right|^2 .$$

Since $Z_{i,n} = g(t_{i,n}) + \varepsilon_{i,n}$, $\forall i = 1,\ldots,n$, and $\widehat{\beta} - \widetilde{\beta} = \widehat{\widetilde{\beta}}$, we obtain

$$\begin{aligned} V_n^2 - \left(1 - \frac{p}{n}\right)\widehat{\sigma}_R^2 = \\ \frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,n}^2 + \sum_{k=1}^{p}\sum_{l=1}^{p}\widehat{\widetilde{\beta}}_k\widehat{\widetilde{\beta}}_l\left(\frac{1}{n}\sum_{i=1}^{n}f_k(t_{i,n})f_l(t_{i,n})\right) - 2\sum_{k=1}^{p}\widehat{\widetilde{\beta}}_k\left(\frac{1}{n}\sum_{i=1}^{n}f_k(t_{i,n})\varepsilon_{i,n}\right) \\ -\frac{1}{2(n-1)}\sum_{i=2}^{n}\Big(g(t_{i,n}) - g(t_{i-1,n})\Big)^2 - \frac{1}{2(n-1)}\sum_{i=2}^{n}\left(\varepsilon_{i,n} - \varepsilon_{i-1,n}\right)^2 \\ -\frac{1}{n-1}\sum_{i=2}^{n}\Big(g(t_{i,n}) - g(t_{i-1,n})\Big)\Big(\varepsilon_{i,n} - \varepsilon_{i-1,n}\Big) + o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

As $t_{i,n},\ i = 1,\ldots,n$ satisfy the condition (A1), we have

$$\frac{1}{n}\sum_{i=1}^{n}f_k(t_{i,n})f_l(t_{i,n}) \;=\; \int_0^1 f_k(t)f_l(t)\,dt + o(1)\,.$$

A straightforward calculation shows

$$\widehat{\widetilde{\beta}}_k\left(\frac{1}{n}\sum_{i=1}^{n}f_k(t_i)\varepsilon_i\right) = o_p\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \widehat{\widetilde{\beta}}_k\widehat{\widetilde{\beta}}_l = o_p\left(\frac{1}{\sqrt{n}}\right),\ k,l = 1,\ldots,p\,.$$

Therefore

$$\widehat{\mathcal{M}}_n^2 - \widetilde{\mathcal{M}}_n^2 = \frac{1}{n}\sum_{i=1}^{n-2}Q_{i,n} - D_n(g) + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{7}$$

where

$$Q_{i,n} = \varepsilon_{i,n}\varepsilon_{i-1,n} + 2\left(g(t_{i,n}) - \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right)\varepsilon_{i,n}$$

and

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

Since the random variables $Q_{i,n}$ are uncorrelated, we have

$$Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n-2}Q_{i,n}\right) = \sigma^4 + 4\sigma^2\frac{1}{n}\sum_{i=1}^{n-2}\left(g(t_{i,n}) - \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right)^2 .$$

Now, using analysis arguments, we show that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n-2}\left(g(t_{i,n}) - \sum_{k=1}^{p}\widetilde{\beta}_k f_k(t_{i,n})\right)^2 = \int_0^1\left(g(t) - \sum_{k=1}^{p}\beta_k f_k(t)\right)^2 dt,$$

where

$$(\beta_1, \ldots, \beta_p)' = \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \int_0^1 \left| g(t) - \sum_{k=1}^{p} \alpha_k f_k(t) \right|^2 dt \right\}.$$

Thus

$$\lim_{n\to\infty} Var\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n-2} Q_{i,n} \right) = \sigma^4 + 4\sigma^2 \mathcal{M}^2(g)\,,$$

where

$$\mathcal{M}^2(g) = \int_0^1 \left| g(t) - \sum_{k=1}^{p} \alpha_k f_k(t) \right|^2 dt.$$

The random variables $Q_i = Q_{i,n},\ i = 1, \ldots, n$ constitute a centered rowwise, 2-dependent array; applying a theorem of Orey (1958), we obtain the result of Theorem 1.

References

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.

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.

Dette, H. and Munk, A., 1998. Validation of linear regression models. *Ann. Stat.*, **26**, 2, 778-800.

Dette, H. and Hetzler, B., 2009. A simple test for the parametric form of the variance function in nonparametric regression. *Ann. Inst. Statist. Math.*, **61**, 4, 861-886.

Dette, H. and Marchlewski, M., 2008. A test for the parametric form of the variance function in a partial linear regression model. *J. Statist. Plann. Inference*, **138**, 10, 3005-3021.

Eubank, R.L. and Hart, J.D., 1992. Testing goodness-of-fit in regression via order selection criteria. *Ann. Stat.*, **20**, 3, 1412-1425.

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.

Härdle, W. and Mammen, E., 1993. Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, **21**, 4, 1926-1947.

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.

Munk, A. and Dette, H., 1998. Nonparametric comparison of several regression functions: exact and asymptotic theory. *Ann. Stat.*, **26**, 6, 2339-2368.

Orey, S., 1958. A central limit theorem for m-dependent random variables. *Duke Math. J.*, **25**, 4, 543-546.

Rice, J., 1984. Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

You, J. and Chen, G., 2005. Testing heteroscedasticity in partially linear regression models. *Statist. Probab. Lett.*, **73**, 1, 61-70.