C. R. Acad. Sci. Paris, Ser. I 349 (2011) 215–217

Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

A nonparametric lack-of-fit test for heteroscedastic regression models

Un test d'adéquation nonparamétrique pour la régression

Jean-Baptiste Aubin, Samuela Leoni-Aubin

INSA Lyon, ICJ, 20, rue Albert-Einstein, 69621 Villeurbanne cedex, France

ARTICLE INFO

Article history:
Received 1 May 2010
Accepted 10 December 2010
Available online 30 December 2010

Presented by Paul Deheuvels

ABSTRACT

A simple test is proposed for examining the correctness of a given completely specified response function against unspecified general alternatives in the context of univariate regression. The usual diagnostic tools based on residual plots are useful but heuristic. We introduce a formal statistical test supplementing the graphical analysis. Technically, the test statistic is the maximum length of the sequences of ordered (with respect to the covariate) observations that are consecutively overestimated or underestimated by the candidate regression function. Note that the testing procedure can cope with heteroscedastic errors and no replicates. Recursive formulae allowing one to calculate the exact distribution of the test statistic under the null hypothesis and under a class of alternative hypotheses are given.

© 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

RÉSUMÉ

Dans le cadre de la régression univariée, nous proposons un outil nonparamétrique général permettant de tester si une fonction connue m est un bon candidat pour la fonction de régression au vu des données. Ce test est basé sur la longueur maximale des suites ordonnées (par rapport à la covariable) des résidus de même signe. Aucune hypothèse n'est faite sur l'homoscédasticité des erreurs. De plus, ce test ne nécessite pas la présence de données répétées. Nous donnons ici la loi de la statistique test sous l'hypothèse nulle que la fonction considérée m est la vraie fonction de régression ainsi que sous une certaine classe d'hypothèses alternatives.

© 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

1. Introduction

Regression is one of the most widely used statistical tools to examine how one variable is related to another. Statisticians usually begin their work by proposing a model for their observations. Then, they have to check on whether this model is correct. The graphical analysis of the residuals is an important step of this process since the detection of a systematic pattern would indicate a misspecified model. Unfortunately, this procedure is heuristic and could lead to errors of interpretation since it is often difficult to determine whether the observed pattern reflects model misspecification or random fluctuations. So it is of interest to complement such an analysis by a formal test. A large literature in this area can be found in Hart [4]. A review of statistical tests and procedures to determine lack of fit associated with the deterministic portion of a proposed linear regression model is presented in Neill and Johnson [7]. We propose a new approach based on maximum length of

E-mail addresses: jean-baptiste.aubin@insa-lyon.fr (J.-B. Aubin), samuela.leoni@insa-lyon.fr (S. Leoni-Aubin).

1631-073X/$ – see front matter © 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crma.2010.12.009

sequences of consecutive overestimated (or underestimated) observations by the model. This test can be computed visually if the sample size is small enough and it is a modification of a nonrandomness test (see Bradley [1, Chap. 11]). In other words, we use it to detect whether residuals are randomly distributed or not.

In Section 2, the length of the longest run test is presented. Section 3 is devoted to the law of the test statistic under the null hypothesis. In Section 4, the power of the test for a class of fixed alternatives is given.

2. The length of the longest run test statistic

Consider a collection of n random variables Y_i generated as $Y_i = m_0(x_i) + \varepsilon_i$, $i = 1, \dots, n$, where the x_i are fixed design points and m_0 is the true regression function. Moreover, the ε_i are independent and centred random variables such that:

$$\forall i = 1, \dots, n, \quad Pr(\varepsilon_i > 0) = Pr(\varepsilon_i < 0) = \frac{1}{2}. \tag{1}$$

Note that no hypothesis is made on the regularity of the function m_0 or on the fact that errors must be identically distributed or homoscedastic, and that normality of ε_i implies Condition (1). *Moreover, contrary to other classical tests* (*like the F-test*), *no replicates are needed to compute our test statistic.*

We address the problem of testing the null hypothesis H_0: $m_0 = m$ vs. H_1: $m_0 \neq m$, where m is a completely specified function.

The i-th residual, $\hat{\varepsilon}_i$, may be seen as substitute for the realisation of the random variable ε_i, thus comprising clues for adequacy or inadequacy of the model assumptions related to the distribution of ε_i. Some classical lack-of-fit test statistics are based on squared residuals, hence their signs are neglected, and we can expect to loose some information. We propose a test statistic that takes these signs into account. This test statistic, L_n, is the maximum length of the sequences of ordered (with respect to the covariate) observations that are consecutively overestimated (or underestimated) by the candidate m. Formally, we define $Z_i := \mathbf{1}_{\{\hat{\varepsilon}_i > 0\}}$, $1 \leqslant i \leqslant n$, $S_0 := 0$, $S_l := Z_1 + \cdots + Z_l$, and put for $0 \leqslant K \leqslant n$, $I^+(n, K) := \max_{0 \leqslant l \leqslant n-K}(S_{l+K} - S_l)$. Let L_n^+ be the largest integer K for which $I^+(n, K) = K$. L_n^+ is the length of the longest run of 1's in $Z_1, \dots, Z_n$, i.e. the length of the longest run of positive residuals. By analogy, we define L_n^- as the length of the longest run of 0's in $Z_1, \dots, Z_n$, that is L_n^- is the largest integer K for which $I^-(n, K) = K$, where $I^-(n, K) := \max_{0 \leqslant l \leqslant n-K}(K - S_{l+K} + S_l)$. Clearly, L_n^- is the length of the longest run of negative residuals. Finally, we define $L_n := \max(L_n^+, L_n^-)$.

For a fixed nominal level $\alpha > 0$, we obtain the following unilateral rejection regions $W_{n,\alpha} = \{L_n > c_{n,\alpha}\}$, where $c_{n,\alpha}$ is the largest integer such that $Pr(L_n > c_{n,\alpha}) \geqslant \alpha$. The corresponding bilateral rejection regions are $W_{n,\alpha}^b = \{L_n \notin [c_{n,1-\alpha/2}, c_{n,\alpha/2}]\}$.

3. Distribution of L_n under the null hypothesis

If m is equal to m_0, then, the residuals $\hat{\varepsilon}_i$ *are* the true errors ε_i. Since Condition (1) holds, we can apply the following recursive formula (Riordan [8, p. 153, Problem 13]): for all $1 < k < n$,

$$\begin{aligned}(n-1)!\,Pr(L_n = k) &= 2(n-2)!\,Pr(L_{n-1} = k) - (n-k-2)!\,Pr(L_{n-k-1} = k) \\ &\quad + (n-2)!\,Pr(L_{n-1} = k-1) - 2(n-3)!\,Pr(L_{n-2} = k-1) + (n-k-1)!\,Pr(L_{n-k} = k-1).\end{aligned}$$

By using $Pr(L_2 = 2) = 1/2$ and $\forall n > 0$, $Pr(L_n = 1) = Pr(L_n = n) = 1/2^{n-1}$, the entire exact distribution of L_n and critical values for every nominal level can be deduced from the above formula.

For most of practical cases of interest, m is estimated. For example, if m is estimated by OLS, an unfortunate property of residuals is that they are autocorrelated even when the true errors are white noise. This divergence from the assumptions disappears in large samples, but may be a problem when performing diagnostic tests in small samples. One way of handling this problem is to transform the OLS residuals so that they do satisfy the LS assumptions when these are correct. One of the most common of these transformations are the so-called *recursive residuals* (see Kianifard and Swallow [5] among others). Another possibility is to estimate m on a subset of the data and to test it on the rest of the data.

In a coin tossing experiment, L_n, L_n^+, and L_n^- can be seen as the length of the longest run of *heads or tails*, *heads* and *tails*, respectively. The length of the longest head run in a coin tossing experiment was investigated in the early days of probability theory. Later, Deheuvels [2] gives upper and lower bounds for L_n^+ for a biased coin.

Schilling [9] discusses the distributions of L_n for unbiased coins, and remarks that for n tosses of a fair coin the length of the longest run of *heads or tails*, statistically speaking, tends to be about one longer than the length of the longest run of *heads* only. For a biased coin, when n is very large, if head is more likely than tail, the distribution function of L_n^+ is well approximated by an extreme value distribution (see Gordon et al. [3]).

4. Distribution of L_n under fixed alternative hypotheses

In this section, we give the distribution of the length of the longest run test statistic under some fixed alternative hypotheses. First of all, we suppose that Condition (1) is fulfilled, and that errors are identically distributed. Moreover, if we test

$$H_0\colon \forall x,\ m_0(x) = m(x) \quad \text{vs.} \quad H_{1,c}\colon \forall x,\ m_0(x) = m(x) + c, \quad c \neq 0,$$

then, under $H_{1,c}$, the probability for an observation to be underestimated (respectively, overestimated), $p(c) \neq \frac{1}{2}$, is constant for all the observations. By considering the total number of positive residuals, k, in the sequence, the cumulative distribution of L_n can be expressed as:

$$P(L_n \leqslant x) = \sum_{k=0}^{n} S_n^{(k)}(x) p(c)^k \bigl(1 - p(c)\bigr)^{n-k},$$

where $S_n^{(k)}(x)$ is the number of sequences of length n that contain k positive residuals in which the length of the longest run of *positive or negative residuals* does not exceed x. Analogously, Schilling [9] studied the cumulative distribution of L_n^+.

In the following proposition, we give a recursive formula to compute the $S_n^{(k)}(x)$.

Proposition 4.1. *Let n and x be such that* $0 < x \leqslant n$. *Then*:

(i) *If* $n - k \leqslant x$ *and* $k \leqslant x$, $S_n^{(k)}(x) = C_n^k$.
(ii) *If* $n - k \leqslant x$ *and* $k > x$, $S_n^{(k)}(x) = \sum_{j=0}^{x} S_{n-j}^{(k)}(x)$.
(iii) *If* $n - k > x$ *and* $k \leqslant x$, $S_n^{(k)}(x) = \sum_{j=0}^{x} S_{n-j}^{(k+1-j)}(x)$.
(iv) *If* $n - k > x$ *and* $k > x$, *let*

$$R_n^{(k)}(x) := \sum_{j \geqslant 0} \left\{ \sum_{i=1}^{x} \bigl\{ S_{n-1-i-2j(x+1)}^{(k-1-j(x+1))}(x) + S_{n-1-i-2j(x+1)}^{(k-i-j(x+1))}(x) - S_{n-1-(2j+1)(x+1)-i}^{(k-(j+1)(x+1))}(x) - S_{n-1-(2j+1)(x+1)-i}^{(k-1-j(x+1)-i)}(x) \bigr\} \right\},$$

with the conventions: $\forall x \in \mathbb{N}^*$ *and* $\forall k \in \mathbb{N}^*$, $S_0^{(0)}(x) := 1$ *and* $S_{-n}^{(-k)}(x) = S_{-n}^{(k)}(x) = S_n^{(-k)}(x) := 0$.

- *If* $\exists (i,j) \in \{1, \ldots, x\} \times \mathbb{N}^*$ *such that* $(k,n) = (2j(x+1) + i, j(x+1))$ *or* $(k,n) = (2j(x+1) + i, j(x+1) + i)$, *then* $S_n^{(k)}(x) = R_n^{(k)}(x) + 1$;
- *if* $\exists (i,j) \in \{1, \ldots, x\} \times \mathbb{N}^*$ *such that* $(k,n) = ((2j+1)(x+1) + i, j(x+1) + i)$ *or* $(k,n) = ((2j+1)(x+1) + i, (j+1)(x+1))$, *then* $S_n^{(k)}(x) = R_n^{(k)}(x) - 1$;
- *else*, $S_n^{(k)}(x) = R_n^{(k)}(x)$.

From this result, one can deduce the exact law of the test statistic under $H_{1,c}$, and the power of the test follows.

In the next proposition, we show that, for n large enough, the distribution function of L_n is well approximated by the distribution function of L_n^+ (or L_n^-, depending on the value of $p(c)$).

Proposition 4.2. *If* $\forall i = 1, \ldots, n$, $Pr(\varepsilon_i > 0) = p(c)$, $p(c) > \frac{1}{2}$ (*resp.* $p(c) < \frac{1}{2}$), *then*

$$\forall k, \quad Pr(L_n \leqslant k) - Pr\bigl(L_n^+ \leqslant k\bigr) = o(1) \quad \textit{when } n \to \infty$$

(*resp.* $Pr(L_n \leqslant k) - Pr(L_n^- \leqslant k) = o(1)$).

The proof is based on Muselli's results [6].

References

[1] J.V. Bradley, Distribution-Free Statistical Tests, Prentice–Hall, 1968.

[2] P. Deheuvels, On the Erdos–Renyi theorem for random fields and sequences and its relationships with the theory of runs and spacings, Z. Wahrsch. Verw. Gebiete 70 (1985) 91–115.

[3] L. Gordon, M.F. Schilling, M.S. Waterman, An extreme value theory for long head runs, Probab. Theory Related Fields 72 (1986) 279–287.

[4] J. Hart, Nonparametric Smoothing and Lack-of-Fit Tests, Springer-Verlag, New York, 1997.

[5] F. Kianifard, W.H. Swallow, A review of the development and application of recursive residuals in linear models, J. Amer. Statist. Assoc. 91 (433) (1996) 391–400.

[6] M. Muselli, Useful inequalities for the longest run distribution, Statist. Probab. Lett. 46 (2000) 239–249.

[7] J.W. Neill, D.E. Johnson, Testing for lack of fit in regression—A review, Comm. Statist. Theory Methods 13 (4) (1984) 485–511.

[8] J. Riordan, An Introduction to Combinatorial Analysis, John Wiley and Sons, 1958.

[9] M.F. Schilling, The longest run of heads, College Math. J. 21 (1990) 196–207.