HAL
open science

A simple lack-of-fit test for a wide class of regression models

Jean-Baptiste Aubin, Samuela Leoni-Aubin

To cite this version:

Jean-Baptiste Aubin, Samuela Leoni-Aubin. A simple lack-of-fit test for a wide class of regression models. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386706

HAL Id: inria-00386706

https://inria.hal.science/inria-00386706

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A SIMPLE LACK-OF-FIT TEST FOR A WIDE CLASS OF REGRESSION MODELS.

Jean-Baptiste AUBIN & Samuela LEONI-AUBIN

jean-baptiste.aubin@utc.fr, U.T.C., Rue P. de Roberval - BP 20529, 60205 Compiègne.
samuela.leoni@insa-lyon.fr, I.N.S.A. de Lyon, 20, Rue A. Einstein, 69621 Villeurbanne.

ABSTRACT

A simple test is proposed for examining the correctness of a given response function against unspecified general alternatives in the context of univariate regression. The usual diagnostic tools based on residuals plots are useful but heuristic. We propose a formal statistical test supplementing the graphical analysis. Technically, the test statistic is the maximum length of the sequences of consecutive (with respect to the covariate) observations overestimated or underestimated by the response function. No hypothesis is made on linearity or on smoothness of the response function, and the testing procedure can also cope with heteroscedastic errors. The exact distribution under the null hypothesis of correctness is obtained. Some results from simulations will be presented at the conference.

RESUME

Un test d'adéquation dans le cas de la régression univariée est proposé. Ce test, très simple, repose sur la longueur maximale d'observations consécutives surestimées (respectivement sous-estimées) par le modèle dont on cherche à tester la qualité. Les principaux avantages de ce test sont son extrème simplicité (on peut calculer sa statistique test visuellement sur des échantillons d'assez petite taille) et sa généralité. En effet, non seulement il ne nécessite pas de faire d'hypothèse sur les fonctions de régression "vraie" et à tester (elles peuvent, par exemple, être discontinues en tout point) mais il permet aussi d'affaiblir des hypothèses faites classiquement sur les erreurs (hétéroscédasticité, non normalité admises). Nous donnons ici la loi exacte de la statistique test. Lors de la conférence, nous présenterons des résultats concernant la puissance de ce test pour une certaine classe d'alternatives, ainsi que des résultats de comparaisons de notre test par rapport à d'autres tests d'adéquation.

MOTS CLES: Modèles de Régression, Choix de Modèles

1. INTRODUCTION

Regression is one of the most widely used statistical tools to examine how one variable is related to another. Regression analysts 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. We propose a new approach based on maximum length of sequences of consecutive overestimated (or underestimated) observations by the model. This omnibus test is very simple and can be computed visually if the sample size is small enough. This test statistic is a modification of a nonrandomness test (see Bradley 1968, chap. 11). In other words, we use this test to detect whether residuals are randomly distributed or not.
In Section 2, we present the form of the proposed test, we give both the unilateral and bilateral rejection regions for levels 0.01, 0.05, and 0.1 for every sample size smaller than 1000.

2. THE TEST

2.1. BASIC MODEL

We consider the following model:

$$Y_i = m(x_i) + \sigma(x_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

where the response variable Y is observed at fixed design points x_i of a covariate x and m is an unknown regression function. Moreover, $\forall i$, $\sigma(x_i) > 0$ and the ε_i are realisations of independent and identically distributed replications of the random variable ε with mean 0, variance 1, and such that:

$$\mathbb{P}(\varepsilon > 0) = \mathbb{P}(\varepsilon < 0) = \frac{1}{2}. \tag{1}$$

Note that no hypothesis is made on the regularity of the function m or on the homoscedasticity of errors, and that normality of ε implies Condition (1). Moreover, contrary to other classical tests (like the F-test), no replicates are needed to compute the Longest Run Test statistic.

We address the problem of testing the null hypothesis

$$H_0 : m = m_0 \qquad vs. \qquad H_1 : m \neq m_0.$$

2.2. THE TEST STATISTIC

When we consider residuals, this may be seen as a substitute for the realisation of vector $\sigma(x_i)\varepsilon_i$, thus comprising clues for adequacy or inadequacy of the model assumptions related to the distribution of $\sigma(x_i)\varepsilon_i$. Some classical lack-of-fit test statistics are based on squared residuals, hence their signs are neglicted, and we can expect to loose some information. We propose a test statistic that takes these signs into account. This test statistics, L_n, is the maximum length of the sequences of consecutive (with respect to

the predictor variable) overestimated (or underestimated) observations. Formally, let $\widehat{\varepsilon}_i$ be the i-th residual, $Z_i := \mathbb{1}_{\{\widehat{\varepsilon}_i > 0\}}$, $1 \leq i \leq n$, $S_0 := 0$, $S_l := Z_1 + \ldots + Z_l$, and put for $0 \leq K \leq n$,

$$I^+(n, K) := \max_{0 \leq l \leq n-K} (S_{l+K} - S_l).$$

Let L_n^+ be the largest integer K for which $I^+(n, K) = K$. L_n^+ is the longest run of 1's in $Z_1, \ldots, Z_n$. By analogy, we define L_n^- as the longest run of 0's in $Z_1, \ldots, Z_n$, that is L_n^- is the largest integer for which

$$I^-(n, K) := \max_{0 \leq l \leq n-K} (K - S_{l+K} + S_l).$$

Finally, we define

$$L_n := \max\left(L_n^+, L_n^-\right).$$

For a fixed nominal level α, 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 $\mathbb{P}(L_n > c_{n,\alpha}) \geq \alpha$.
One can also detect if L_n is too small by performing a bilateral test taking

$$W_{n,\alpha} = \left\{L_n \notin [c_{n,\alpha/2}, c_{n,1-\alpha/2}]\right\}.$$

2.3. LAW OF L_n UNDER THE NULL HYPOTHESIS

If m is completely specified (i.e., it contains no unknown parameters to be estimated) and is equal to m_0, then, the residuals $\widehat{\varepsilon}_i$ are the true errors ε_i. Since Condition 1 holds, we can apply to L_n the following recursive formula from Riordan (1958):

$$\begin{aligned}(n-1)!\ \mathbb{P}(L_n = k) \;=\; & 2(n-2)!\ \mathbb{P}(L_{n-1} = k) - (n-k-2)!\ \mathbb{P}(L_{n-k-1} = k) \\ & +(n-2)!\ \mathbb{P}(L_{n-1} = k-1) - 2(n-3)!\ \mathbb{P}(L_{n-2} = k-1) \\ & +(n-k-1)!\ \mathbb{P}(L_{n-k} = k-1).\end{aligned} \tag{2}$$

By using $\mathbb{P}(L_2 = 2) = 1/2$ and $\forall\ n > 0$, $\mathbb{P}(L_n = 1) = 1/2^{n-1}$, the entire exact law of L_n can be deduced from the above formula.
For most of practical cases of interest m is estimated. Nevertheless, if m is consistently estimated, then, for n large enough, Formula 2 still holds.
Schilling (1990) discusses the distributions of L_n for biased and unbiased coins, and remarks that for n tosses of a fair coin the longest run of *heads or tails*, statistically speaking, tends to be about one longer than 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 the distribution function of L_n^+.
For every n smaller than 1000, we give exact critical values c_α with the corresponding α levels for the unilateral and bilateral tests, respectively. α is taken successively equal to 0.01, 0.05, and 0.1 (see Tables 1 and 2 on pages 4 and 5).

n	$c_{n,0.1}$	$c_{n,0.05}$	$c_{n,0.01}$
5	4	–	–
6	4	5	–
7	5	5	–
8	5	6	7
9 – 10	5	6	8
11	6	6	8
12	6	7	8
13 – 17	6	7	9
18	7	7	9
19 – 29	7	8	10
30 – 32	7	8	11
33	7	9	11
34 – 51	8	9	11
52 – 60	8	9	12
61 – 93	9	10	12
94 – 113	9	10	13
114	9	11	13
115 – 176	10	11	13
177 – 219	10	11	14
220 – 223	10	12	14
224 – 343	11	12	14
344 – 430	11	12	15
431 – 442	11	13	15
443 – 673	12	13	15
674 – 851	12	13	16
852 – 872	12	14	16
873 – 1000	13	14	16

Table 1: Critical values $c_{n,\alpha}$ for different levels α and sample sizes n of the unilateral test.

3. CONCLUSIONS AND PERSPECTIVES

We consider the lack-of-fit testing problem by using a nonparametric test based on a simple geometrical idea.

n	$c_{n,0.005}$	$c_{n,0.995}$	n	$c_{n,0.025}$	$c_{n,0.975}$	n	$c_{n,0.05}$	$c_{n,0.95}$
9	–	9	7 – 8	–	7	6 – 7	–	6
10 – 13	–	10	9 – 12	–	8	8 – 11	–	7
14 – 19	–	11	13 – 19	–	9	12 – 15	–	8
20 – 26	–	12	20 – 33	–	10	16 – 18	–	8
27 – 30	–	12	34 – 46	–	11	19 – 32	–	9
31 – 52	–	13	47 – 60	–	11	33 – 38	–	10
53 – 65	–	14	61 – 103	–	12	39 – 60	–	10
66 – 94	–	14	104 – 113	2	12	61 – 84	–	11
95 – 146	–	15	114 – 218	2	13	85 – 113	2	11
147 – 177	2	15	219	2	14	113 – 178	2	12
178 – 312	2	16	220 – 426	3	14	179 – 219	3	12
313 – 342	3	16	427 – 452	3	15	220 – 368	3	13
343 – 647	3	17	453 – 842	4	15	369 – 430	4	13
648 – 671	4	17	843 – 922	4	16	431 – 750	4	14
672 – 1000	4	18	923 – 1000	5	16	751 – 851	5	14
						852 – 1000	5	15

Table 2: Critical values $c_{n,\alpha}$ and $c_{n,(1-\alpha)}$ for different levels α and sample sizes n of the bilateral test.

The required computation for our method can be accomplished by using the standard statistical software packages and even visually for small samples.
Note that the Longest Run Test can be used in more general frameworks as heteroscedasticity, non-normality of errors (which is not the case for most of the others), or without hypothesis on smoothness of m. Nevertheless, the lost of information due to its extreme simplicity should bring the user to see it as a useful complement to other more classical tests.
An extension of this work can be the application of this test in the context of multiple regression. This development will be reported in the future.
It appears from the simulations (that will be presented at the conference) that our simple and general test can be a serious concurrent to other tests even in a specific framework as linear regression.

BIBLIOGRAPHY

Bradley, J. V. (1968). *Distribution-Free Statistical Tests.* Prentice-Hall Inc.

Riordan, J. (1958). *An Introduction to Combinatorial Analysis.* John Wiley and sons, Inc.

Schilling, M. F. (1990). The Longest Run of Heads. *College Math. J.* **21**:196–207.