ELSEVIER

Computational Statistics & Data Analysis 35 (2000) 211–231

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Resampling for checking linear regression models via non-parametric regression estimation

J.M. Vilar Fernández[a,1] , W. González Manteiga[b,*,2]

[a] *Departamento de Matemáticas, Facultad de Informática, Universidad de A Coruña 15071, Spain*
[b] *Departamento de Estadística e I.O., Facultad de Matemáticas, Universidad de Santiago 15706, Spain*

Received 1 June 1998; received in revised form 1 November 1999; accepted 26 November 1999

Abstract

Let us consider the fixed regression model, $Y_t = m(x_t) + \varepsilon_t,\ t = 1,\dots,n$, and assume that the random errors, $\{\varepsilon_t\}$, follow an ARMA-type dependence structure. The purpose of this paper is to study the application of the bootstrap test to check that the unknown regression function, m, follows a general linear model of the type:

$$\mathrm{H}_0\colon m \in \mathscr{M} = \{m_\theta(\cdot) = A^t(\cdot)\theta\colon \theta \in \Theta \subset \mathbb{R}^q\}$$

with A being a functional of $\mathbb{R}$ in $\mathbb{R}^q$. In a previous paper, González-Manteiga and Vilar-Fernández (1995) proposed a test, $D = d^2(\hat{m}_n, m_{\hat{\theta}_n})$, based on the Crámer–von-Mises-type functional distance, where $\hat{m}_n$ is a Gasser–Müller-type non-parametric estimator of m, and $m_{\hat{\theta}_n}$ is a member of the family $\mathscr{M}$ that is closest to $\hat{m}_n$. In this work, two bootstrap algorithms are considered, where the dependence structure of the errors is taken into account. A broad simulation study and an applied example show the good behavior of the bootstrap test. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Hypothesis testing; Non-parametric estimators; Regression models; Time series; Bootstrap

1. Introduction

Let us consider the regression model

$$Y_t = m(x_t) + \varepsilon_t \quad (t = 1,\dots,n), \tag{1}$$

* Corresponding author. Tel.: +34-981-563100e13204; fax: +34-981-597054.
E-mail address: wences@zmat.usc.es (W.González Manteiga).

[1] Partially supported by the grants PB94-0494 and XUGA10501B97.
[2] Partially supported by the grants PB95-0826 and XUGA20701B96.

PII: S 0167-9473(99)00117-6

where $x_t \in C$, with $\mathbb{C}$ a compact set in $\mathbb{R}$, $m(\cdot)$ the unknown regression function, and $\{\varepsilon_t\}_{t=1}^n$ a sequence of unobserved zero mean random variables.

In the last few years, several hypothesis tests have been developed for testing

$$H_0\colon m \in \mathscr{M} = \{m_\theta(\cdot)\colon \theta \in \Theta \subset \mathbb{R}^q\} \tag{2}$$

versus a general alternative hypotheses of the type:

H_1: "m is a function with a certain degree of smoothness".

Given an initial sample $\{(x_t, Y_t)\}_{t=1}^n$, almost all of these tests are based on one distance between a non-parametric pilot estimator, $\hat{m}_n$, and one parametric estimator, $m_{\hat{\theta}}$, of m under H_0, denoted by $D = d^2(\hat{m}_n, m_{\hat{\theta}_n})$. If this discrepancy is statistically significant, hypothesis H_0 is rejected. Otherwise, it is accepted. Among the interesting recent papers that address this problem we can cite those by Firth et al. (1991), Kozek (1991), Eubank and Hart (1993), Eubank et al. (1993), Härdle and Mammen (1993) and Stute and González-Manteiga (1996), where different non-parametric pilot estimators are used (kernel, spline, etc.).

In this work, we devote attention to the goodness of fit for linear regression models. That is, for type (1) models we wish to test the hypothesis

$$H_0\colon m \in \mathscr{M}_1 = \{m_\theta(\cdot) = A^t(\cdot)\theta\colon \theta \in \Theta \subset \mathbb{R}^q\} \tag{3}$$

with respect to the alternative given in (2), where A is a functional of $\mathbb{R}$ of $\mathbb{R}^q$.

The study is carried out taking into account that the errors, ε_t, are dependent. It often happens when analyzing economical data samples, growth curves and, in general, whenever the observations are sequentially gathered in time. It is important to take into account the existence of the correlation among the errors when the model is statistically analyzed. To ignore this fact causes inefficiency in the parametric estimation of the model (Seber and Wild, 1989), in the non-parametric estimation (Chu and Marron, 1991), and it also affects the power of the goodness-of-fit test used, as we will later show in the simulation study.

The dependence structure in the errors for the goodness-of-fit problem, was considered for the first time in a previous paper by González-Manteiga and Vilar-Fernández (1995). In their work, a test based on a discrepancy between a non-parametric estimator of m (of Gasser and Müller, 1979 type)

$$\hat{m}_n(x) = \sum_{j=1}^n h^{-1}\left\{\int_{s_{j-1}}^{s_j} K\left(\frac{x-s}{h}\right) \mathrm{d}s\right\} Y_j = \sum_{j=1}^n W_{nj}(x) Y_j \tag{4}$$

and one parametric, $m_{\hat{\theta}}$, was considered. Denoting as usual $C = [0,1]$, $s_0 = 0$, $s_{j-1} \le x_j \le s_j$, $s_n = 1$, $j = 1, \dots, n$, K a kernel function, and $h > 0$ the smoothing parameter. The estimator $\hat{\theta}$ minimizes the functional

$$\Psi(\theta) = \int (\hat{m}_n(x) - A^t(x)\theta)^2 \omega(x)\, \mathrm{d}\Omega_n(x) = \frac{1}{n}\sum_{i=1}^n (\hat{m}_n(x_i) - A^t(x_i)\theta)^2 \omega(x_i), \tag{5}$$

where ω is a weight function in order to prevent boundary effects of the kernel estimator and Ω_n is the empirical distribution function over the points of the design.