Computational Statistics and Data Analysis 72 (2014) 45–56

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Testing constancy in monotone response models

Ana Colubi, J. Santos Domínguez-Menchero*, Gil González-Rodríguez

Dpto. de Estadística. Universidad de Oviedo. 33007 Oviedo, Spain

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form 31 October 2013
Accepted 31 October 2013
Available online 8 November 2013

Keywords:
Isotonic regression
Order restricted statistical inference
Chi-bar-squared tests
Bootstrap test
Independent and dependent data

ABSTRACT

A model in which the response is monotonically related to a given exposure or predictor is considered. This is motivated by dose–response analysis, however it also applies to survival distributions depending on a series of ordered multinomial parameters or, in a more general context, to change-point problems. In these contexts, although the monotonicity of the response may be *a priori* known, it is often crucial to determine whether the relationship is effective in a given interval, in the sense of not being constant. An efficient nonparametric test for the constancy of the regression when it is known to be monotone is developed for both independent and dependent data. The asymptotic null distribution of a test statistic based on the integrated regression function is obtained. The power against local alternatives is investigated, and the improvements with respect to the previous studies in the topic are shown. Some bootstrap procedures for the case of independent and dependent data are developed and employed in several applications.

1. Introduction

The relationship between a given response and a predictor is sometimes known to be monotone in some way. In biomedical studies examples concerning growth curves, dose–response models, disease risks-biomarkers analysis etc. are found (see Bornkamp and Ickstadt, 2009, Ghosh, 2007, Marschner et al., 2012, Salanti and Ulm, 2003). The monotonicity of the regression function can be tested in various ways (see, for instance, Robertson et al., 1988, Ghosal et al., 2000, Hall and Heckman, 2000, Gijbels et al., 2000, Dumbgen and Spokoiny, 2001, Domínguez-Menchero et al., 2005 or Meyer, 2008, and the references therein). Some other papers (see, for instance, Delgado and Escanciano, 2012) allow for testing the monotonicity of the conditional distributions and its moments.

The consideration of the shape constraint in the estimation problems of monotone relationships obviously provides more meaningful conclusions (see Banerjee, 2007, Bhattacharya and Lin, 2013, Gunn and Dunson, 2005, Hazelton and Turlach, 2011). As a complement of those kinds of studies it raises the question of determining whether the monotone relationship is effective in a given interval. In this sense, the relationship between a disease risk and the distance to the point source has been proposed to be (non-parametrically) estimated under the restriction that the risk should be non-increasing with distance (see Diggle et al., 1999). An essential question is to verify whether within a given radius the risk effectively decreases or, on the contrary, it is constant irrespective of the distance to the prespecified point. In the same way, one can consider the response at low dose levels (see Neelon and Dunson, 2004) or the constancy of ordered multinomial parameters in survival distributions, as the proportion of menopause with the age (see Jewell and Kalbfleisch, 2004), to name but a few.

These problems can generally be stated in terms of a test for the constancy of the isotonic regression (see, for instance, Robertson et al., 1988, Brillinger, 1989, Wu et al., 2001 and Colubi et al., 2006, 2007). The results in Robertson et al. (1988) and in Colubi et al. (2006, 2007) are concerned with independent error models from different points of view. The results in Wu et al. (2001) extend those in Robertson et al. (1988) to the non-independent data case and improve those in Brillinger

* Corresponding author.
E-mail addresses: colubi@uniovi.es (A. Colubi), jsdm@uniovi.es (J.S. Domínguez-Menchero), gil@uniovi.es (G. González-Rodríguez).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.10.029

(1989) in the same context. They are based on some modifications of the classical isotonic likelihood ratio test in Robertson et al. (1988) obtained by maximum likelihood of a penalized function.

In many situations goodness of fit tests can be used for testing constancy of the regression function. In this respect Durot (2008) and Durot and Reboul (2010) proposed a goodness of fit test to a parametric family of isotonic regression functions on a random design setting. It should be noted that the mentioned goodness of fit tests are not applicable for testing constancy as the regression functions in the parametric class cannot have null derivative at any point, which discards the family of constant regression functions.

In order to cope efficiently with most of the practical situations a new way of testing the constancy of an isotonic regression for dependent data is proposed and analysed here. The analysis of local alternatives reveals an improvement w.r.t. the previous studies. The enhancement is also empirically illustrated. The proposed procedure is based on Stute (1997) and Stute et al. (1998), where the integrated regression function is used for testing the constancy of a general regression function without restricting its shape for random design and independent observations. A related strategy is also used in Durot (2003) for testing isotonicity versus a non-restricted shape for fixed design and independent observations. Stute's procedure is specialized here for testing constancy against an isotonic behavior of the regression function, including both fixed and random design as well as the case of dependent observations. Moreover, the obtained results also allow the extension of the test in Durot (2003) for random design and dependent cases.

The rest of the paper is organized as follows. In Section 2 the model and the testing problem are motivated. In Section 3 the theoretical results supporting the bootstrap testing procedures, as well as comparisons to other approaches are included (the proofs are left for the Appendix). In Section 4 the approach is applied to some case-studies.

2. Monotone response models and testing problem

Monotone response models refer to a wide class of situations in which the conditional distribution of a response given a predictor depends on a parameter which is a monotone function of the predictor (see Banerjee, 2007). It may be assumed that the conditional distribution comes from a given parametric class, although in this paper we will consider a more general non-parametric situation. The most common response models refer to the conditional mean, such as those concerning binary choice or Poisson regression models. The first ones are useful, for instance, to evaluate the probability of disease given the level exposure to a toxin or the value of a biomarker (see Ghosh, 2007, Hunt and Rai, 2005). Also in survival analysis a response such as the menopausal status is explained as a function of the age (see Jewell and Kalbfleisch, 2004). The second ones are related to problems such as the above-mentioned disease risk estimation depending on the distance to a point source (see Diggle et al., 1999).

The monotone regression model is usually written as $Y(x_{i,n}) = m(x_{i,n}) + \varepsilon(x_{i,n})$ for a fixed design $\{x_{1,n}, \dots, x_{n,n}\} \subset A$, with $x_{i,n} < x_{j,n}$ $(1 \le i < j \le n)$ in an interval $A \subseteq \mathbb{R}$. The conditional mean response $m(x)$ and the conditional variance $\sigma^2(x)$ is finite for all $x \in A$, and the regression function m is known to be isotonic. Under some conditions, conclusions for this model can be also extended to a random design.

In fact, Y, m and ε may depend also on n, although we do not make it explicit in the notation for the sake of clearness. In this way, the model includes time series $Y_i = m_i + \varepsilon_i$, $i = 1, \dots$ by taking, for instance, $A = [0, 1]$, $x_{i,n} = i/(n+1)$, and $m_n(x_{i,n})$, $\varepsilon_n(x_{i,n})$, $Y_n(x_{i,n})$ as m_i, ε_i and Y_i respectively.

For the binary choice models we have a dichotomous response variable ($Y = 1$ or $Y = 0$) and a continuous predictor X such that $P(Y = 1|X) = m(X)$, and the link function m is often known to be monotone. For Poisson regression models the response Y is assumed to follow a Poisson distribution whose mean m continuously depends on an exposure X, that is, $Y(x) \sim \mathcal{P}(m(x))$. In cases such as spatial disease risks, m is assumed to be non-increasing.

The monotone regression function m may be piecewise constant, that is, constant in a given interval. Testing this hypothesis is frequently critical, as it implies that the mean response is the same irrespectively of the levels of exposure in such an interval. For this reason, we pose the test

$$H_0 : m \text{ is constant} \quad \text{versus} \quad H_1 : m \text{ is not constant} \tag{1}$$

under the monotonicity assumption of m. The sample information consists of $r_{i,n}$ observations of $Y(x_{i,n})$, that is, a set $\{Y^j(x_{i,n})\}_{j=1}^{r_{i,n}}$ with $Y^j(x_{i,n}) = m(x_{i,n}) + \varepsilon^j(x_{i,n})$ $(i = 1, \dots, n)$.

3. The new testing procedure based on the integrated regression function

Let F be a continuous distribution function on A. It is well known that the Lebesgue integral regression function, or integrated regression function, $M(t) = \int_0^t m(F^{-1}(\tau))d\tau$ determines the function m under mild conditions (see, for instance Stute et al., 1998).

Let F_n be the empirical distribution of the design points $x_{1,n}, \dots, x_{n,n}$. If $F_n \overset{n\to\infty}{\longrightarrow} F$, a uniform consistent estimator of M can be established as follows. Consider the Cumulative Sum Diagram of a given real function f, that is, the mapping defined on each $t \in [0, 1]$ as the linear interpolation from the values $\textsc{Csd}_f(0) = 0$ and $\textsc{Csd}_f(i/n) = \sum_{k=1}^{i} f(x_{k,n})/n$ for $i = 1, \dots, n$. Then, M can be estimated by $\textsc{Csd}_{\hat{m}}$, where $\hat{m}$ is so that $\hat{m}(x_{i,n}) = \sum_{j=1}^{r_{i,n}} \frac{Y^j(x_{i,n})}{r_{i,n}}$, $i = 1, \dots, n$.