Computational Statistics and Data Analysis 97 (2016) 151–168

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust testing for superiority between two regression curves

Graciela Boente [a], Juan Carlos Pardo-Fernández [b,*]

[a] Departamento de Matemáticas, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428, Buenos Aires, Argentina

[b] Departamento de Estatística e Investigación Operativa, Universidade de Vigo, Campus Universitario As Lagoas-Marcosende, Vigo, 36310, Spain

ARTICLE INFO

Article history:
Received 23 June 2015
Received in revised form 30 November 2015
Accepted 1 December 2015
Available online 14 December 2015

Keywords:
Hypothesis testing
Nonparametric regression models
Robust inference
Smoothing techniques

ABSTRACT

The problem of testing the null hypothesis that the regression functions of two populations are equal versus one-sided alternatives under a general nonparametric homoscedastic regression model is considered. To protect against atypical observations, the test statistic is based on the residuals obtained by using a robust estimate for the regression function under the null hypothesis. The asymptotic distribution of the test statistic is studied under the null hypothesis and under root−n local alternatives. A Monte Carlo study is performed to compare the finite sample behaviour of the proposed tests with the classical one obtained using local averages. A sensitivity analysis is carried on a real data set.

1. Introduction

Let us assume that the random vectors $(X_j, Y_j)^{\mathrm{T}} \in \mathbb{R}^2, j = 1, 2$, follow the homoscedastic nonparametric regression models given by

$$Y_j = m_j(X_j) + \varepsilon_j = m_j(X_j) + \sigma_j U_j, \tag{1}$$

where $m_j : \mathbb{R} \to \mathbb{R}$ is a nonparametric smooth function and the error ε_j is independent of the covariate X_j. Throughout this paper, we will not require any moment conditions on the error distributions. As is usual in a robust framework, let us assume that the errors ε_j are such that $\varepsilon_j = \sigma_j U_j$, where U_j has a symmetric distribution $G_j(\cdot)$ with scale 1, so that we are able to identify the error's scale, σ_j. When second moments exist, as the case of the classical approach is, these conditions imply that $\mathbb{E}(\varepsilon_j) = 0$ and $\mathrm{VAR}(\varepsilon_j) = \sigma_j^2$, which means that m_j represents the conditional mean, while σ_j^2 equals the residuals variance, i.e., $\sigma_j^2 = \mathrm{VAR}(Y_j - m_j(X_j))$. The nonparametric nature of model (1) offers more flexibility than the standard linear model when modelling a complicated relationship between the response variable and the covariate. In many situations, it is of interest to compare the regression functions m_1 and m_2 to decide if the same functional form appears in both populations. In particular, in this paper we focus on testing the null hypothesis of equality of the regression curves versus a one-sided alternative. Let $\mathcal{R}$ be the common support of the covariates X_1 and X_2 where the comparison will be performed. The null hypothesis to be considered is

$$\mathrm{H}_0 : m_1(x) = m_2(x) \quad \text{for all } x \in \mathcal{R},$$

* Corresponding author. Fax: +34 986 812 401.
E-mail addresses: gboente@dm.uba.ar (G. Boente), juancp@uvigo.es (J.C. Pardo-Fernández).

http://dx.doi.org/10.1016/j.csda.2015.12.002
0167-9473/

while the alternative hypothesis is of the following one-sided type

$$\begin{aligned} &H_1 : m_1(x) \le m_2(x) \text{ for all } x \in \mathcal{R} \text{ and } m_1(x) < m_2(x) \text{ for } x \in \mathcal{A}, \\ &\quad \text{where } \mathcal{A} \subset \mathcal{R} \text{ is such that } \mathbb{P}(X_j \in \mathcal{A}) > 0, \ \text{for } j = 1, 2. \end{aligned} \tag{2}$$

When second moments exist, the problem of testing equality of two regression curves versus one-sided alternatives has been considered by several authors such as Hall et al. (1997), Koul and Schick (1997, 2003) and Neumeyer and Dette (2005), who extended the test proposed in Speckman et al. (2003) to allow for heteroscedasticity. On the other hand, Neumeyer and Pardo-Fernández (2009) introduced a simple root-n test statistic based on the comparison of the sample averages of the estimated residuals, which were computed with respect to a linear convex combination of the kernel regression estimators obtained from each sample.

As is well known, linear kernel regression estimators are sensitive to atypical observations, since they are based on averaging the responses. When estimating the regression function at a value x, the effect of an outlier in the responses will be larger as the distance between the related covariate and the point x is smaller. In this sense, atypical data in the responses in nonparametric regression may lead to a complete distorted estimation which will clearly influence the test statistic and the conclusions of the testing procedure. In this sense, robust estimates are needed in order to provide more reliable estimations and inferences. Beyond the importance of developing robust estimators, the problem of obtaining robust hypothesis testing procedures also deserves attention. In linear regression, recent developments were given, among others, by Salibian-Barrera et al. (2016), where also references to previous robust proposals can be found. However, in the nonparametric setting, robust testing procedures are very scarce. Recently, Dette and Marchlewski (2010) considered a robust test for homoscedasticity in nonparametric regression. On the other hand, under a partly linear regression model, Bianco et al. (2006) proposed a test to study if the nonparametric component equals a fixed given function, while Boente et al. (2013) considered the hypothesis that the nonparametric function is a linear function under a generalized partially linear model. For the problem of testing superiority between two regression curves, Koul and Schick (1997) defined a family of covariate-matched statistics and derived its asymptotic behaviour under the null hypothesis and under root-n local alternatives. This family includes, in particular, a covariate-matched Wilcoxon–Mann–Whitney test based on the sign of all response differences which does not require the existence of second moments. Besides, these authors provide an asymptotic optimality theory allowing to obtain locally asymptotically minimax tests against nonparametric root-n alternatives. To derive these properties, Koul and Schick (1997) assume equal error distributions and equal design densities. In order to avoid these assumptions, Koul and Schick (2003) developed a modified version of one of the covariate-matched statistics based on the response differences of Koul and Schick (1997), but this statistic is not robust when atypical data arise in the responses, as it assumes the existence of second moments. When considering the problem of comparing two or more regression functions, Feng et al. (2015) considered a test for H_0 versus the general alternative $m_1 \neq m_2$ using a generalized likelihood ratio test incorporating a Wilcoxon likelihood function and kernel smoothers, which allows to detect alternatives with rate $\sqrt{nh}$, where h is the bandwidth parameter; however, these authors assume the existence of second moment of the regression errors, so the applicability of their method in a robust context is quite limited.

The aim of this paper is to propose a class of robust tests for H_0 versus H_1 in (2) which allows for possibly different covariate densities and error densities in the two populations. Our proposal combines the ideas of robust smoothing with those given in Neumeyer and Pardo-Fernández (2009) to obtain a procedure detecting root-n alternatives. In Section 2, we recall the definition of the robust estimators. The test statistics is introduced in Section 3, where its asymptotic behaviour under the null hypothesis and root-n local alternatives is also studied. We present the results of a Monte Carlo study in Section 4 and an illustration to a real data set in Section 5. The Appendix A contains some auxiliary results about the robust nonparametric estimator presented in Section 2 and the proof of our main result.

2. Basic definitions and notation

Throughout this paper, we consider independent and identically distributed observations $(X_{ij}, Y_{ij})^{\mathrm{T}}$, $1 \le i \le n_j$, with the same distribution as $(X_j, Y_j)^{\mathrm{T}}$, $j = 1, 2$. When $\mathbb{E}|Y_j| < \infty$, the regression functions m_j in (1), which in this case equals $\mathbb{E}(Y_j|X_j)$, can be estimated by using the Nadaraya–Watson estimator (see, for example Härdle, 1990). To be more precise, let K be a kernel function (usually a symmetric density) and $h = h_n$ a sequence of strictly positive real numbers. Denote $K_h(u) = h^{-1}K(u/h)$. Then, the classical regression estimators of m_j are defined as

$$\widehat{m}_{j,\mathrm{CL}}(x) = \left\{ \sum_{\ell=1}^{n_j} K_h\left(x - X_{\ell j}\right) \right\}^{-1} \sum_{i=1}^{n_j} K_h\left(x - X_{ij}\right) Y_{ij}. \tag{3}$$

As mentioned in the introduction, the estimators defined in (3) are sensitive to atypical observations, since they are based on averaging the responses. Robust estimates in a non-parametric setting need to be employed to provide estimators insensitive to a single wild spike outlier. Several proposals have been considered and studied in the literature. We can mention, among others, Härdle and Tsybakov (1988) and Boente and Fraiman (1989), who considered robust equivariant estimators under a general heteroscedastic regression model. It is well known that, under a homoscedastic regression model, root-n scale estimators can be obtained. In particular, for fixed designs, scale estimators based on differences are widely