Journal of Statistical Planning and Inference 139 (2009) 2836 – 2850

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Goodness-of-fit tests for parametric regression with selection biased data

Jorge L. Ojeda Cabrera[a], Ingrid Van Keilegom[b,*]

[a] Fac. de Ciencias, Edif. Matemáticas, Universidade de Zaragoza, Pedro Cerbuna, num. 12, 50009 Zaragoza, Spain
[b] Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 24 January 2008
Received in revised form
31 October 2008
Accepted 10 January 2009
Available online 21 January 2009

Keywords:
Biased sampling
Bootstrap
Empirical process
Goodness-of-fit test
Heteroscedastic model
Location-scale regression
Model diagnostics
Nonparametric regression
Weak convergence

ABSTRACT

Consider the nonparametric location-scale regression model $Y = m(X) + \sigma(X)\varepsilon$, where the error ε is independent of the covariate X, and m and σ are smooth but unknown functions. The pair (X, Y) is allowed to be subject to selection bias. We construct tests for the hypothesis that $m(\cdot)$ belongs to some parametric family of regression functions. The proposed tests compare the nonparametric maximum likelihood estimator (NPMLE) based on the residuals obtained under the assumed parametric model, with the NPMLE based on the residuals obtained without using the parametric model assumption. The asymptotic distribution of the test statistics is obtained. A bootstrap procedure is proposed to approximate the critical values of the tests. Finally, the finite sample performance of the proposed tests is studied in a simulation study, and the developed tests are applied on environmental data.

1. Introduction

Consider the nonparametric location-scale regression model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{1}$$

where Y is the variable of interest, X is a covariate, the error ε is independent of X, $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = 1$, and $m(\cdot)$ and $\sigma^2(\cdot)$ are smooth but unknown regression and variance curves, respectively.

Suppose that we cannot observe the pair (X, Y) directly, but that our sample comes from (X^w, Y^w), a bivariate random vector whose distribution is given by

$$\mathrm{d}F^w(x,y) = \frac{w(x,y)}{\mu_w}\mathrm{d}F(x,y), \tag{2}$$

where $F(x,y)$ is the bivariate distribution of (X, Y), and μ_w is the mean value of $w(X, Y)$. The weight function $w(X, Y)$ drives the relationship between the observed random vector (X^w, Y^w), and the unobserved random vector (X, Y). Let $(X_1^w, Y_1^w), \ldots, (X_n^w, Y_n^w)$ be n independent replications of (X^w, Y^w).

A nice description of practical situations where selection bias is encountered, can be found in Rao (1997). While in Rao (1997) discrete populations are considered, Mahfoud and Patil (1982) and Patil and Taillie (1989) also deal with the continuous and multivariate framework (e.g. $w(x,y) = y^\alpha$, $w(x,y) = \max(x,y)$, $w(x,y) = \min(x,y)$, $w(x,y) = x + y$). A more theoretical treatment of the problem can be found in a series of papers by Vardi (1982). In that paper, the nonparametric maximum likelihood estimator

* Corresponding author.
E-mail addresses: jojeda@unizar.es (J.L. Ojeda Cabrera), ingrid.vankeilegom@uclouvain.be (I. Van Keilegom).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.01.008

(NPMLE) of the distribution function is obtained for the special case where the data are univariate and are subject to length-bias sampling (i.e. when $w(x,y)=y$). See also Vardi (1985) and Gill et al. (1988), where the general case of selection biased sampling is considered for univariate data. They also develop a number of examples (including stratification and length-bias) and give a detailed treatment of the large sample properties of the NPMLE. The problem of estimating the density from biased data has been considered by Efromovich (2004), while the paper by Cristóbal and Alcalá (2001) is a valuable source of references on nonparametric regression function estimation from biased data.

The aim of this paper is twofold. We will first develop the asymptotic properties of the NPMLE of the error distribution, based on the nonparametric residuals $\hat{\varepsilon}_i^w=\{Y_i^w-\hat{m}_n(X_i^w)\}/\hat{\sigma}_n(X_i^w)$ $(i=1,\dots,n)$, where $\hat{m}_n(\cdot)$ and $\hat{\sigma}_n(\cdot)$ are appropriate kernel estimators of $m(\cdot)$ and $\sigma(\cdot)$, respectively. The proofs of these properties rely heavily on modern empirical process techniques, needed to take care of the difference $I(\hat{\varepsilon}_i^w\leqslant y)-I(\varepsilon_i^w\leqslant y)$ $(i=1,\dots,n)$, where $\varepsilon_i^w=\{Y_i^w-m(X_i^w)\}/\sigma(X_i^w)$. See also Akritas and Van Keilegom (2001) and Van Keilegom and Akritas (1999), where the NPMLE of the error distribution has been studied for completely observed and right censored observations, respectively. Related estimation problems can be found in Cheng (2004), Efromovich (2005) and Müller et al. (2004a, b).

Secondly, we will develop appropriate test statistics for the hypothesis

$$\mathrm{H}_0\colon m(\cdot)\in\mathscr{M}, \tag{3}$$

when the data are subject to selection bias. Here, $\mathscr{M}=\{m_\theta(\cdot):\theta\in\Theta\}$ is a class of parametric regression functions, including (but not restricted to) the class of linear regression functions. The set Θ is supposed to be a closed subset of $\mathbb{R}^d$. We are interested in developing omnibus tests, which have power against any alternative hypothesis.

The test statistics proposed in this paper are based on the following idea. When the null hypothesis H_0 is true, then the NPMLE based on the 'parametric' residuals $\hat{\varepsilon}_{0i}^w=\{Y_i^w-m_{\hat{\theta}}(X_i^w)\}/\hat{\sigma}_n(X_i^w)$, where $\hat{\theta}$ is an appropriate estimator of θ under H_0, will be close to the NPMLE based on the nonparametric residuals $\hat{\varepsilon}_i^w$ $(i=1,\dots,n)$. It is worth noting that when the data suffer from selection-bias, not only the sampled data are biased, but also the residuals. The two estimators will be compared through Kolmogorov–Smirnov and Cramér–von Mises type statistics. Similar test statistics have been considered by Van Keilegom et al. (2008) and Pardo-Fernández et al. (2007a) for directly observed and right censored observations, respectively. See also Neumeyer et al. (2006), Pardo-Fernández and Van Keilegom (2006), Einmahl and Van Keilegom (2008) and Pardo-Fernández et al. (2007b) for other testing procedures under model (1). However, the above papers use local constant smoothing, whereas in this paper we will use local linear smoothing, because of the well-known advantages in terms of asymptotic bias of estimators based on local linear instead of local constant smoothing.

The paper is organized as follows. In the next section, we propose an estimator of the error distribution, and study its asymptotic properties. Section 3 is devoted to the construction and study of test statistics for the hypothesis (3). A bootstrap approximation is defined in Section 4, which is a useful alternative for the normal approximations obtained in Sections 2 and 3. Some simulation results are summarized in Section 5, while Section 6 shows the results of the analysis of data on the acid neutralizing capacity of lakes in the Northeastern states of the US. Finally, the proofs of the asymptotic results are given in Appendix A.

2. Estimation of the error distribution

We start with introducing a number of notations. Let $F_\varepsilon(e)=P(\varepsilon\leqslant e)$ and $F_X(x)=P(X\leqslant x)$. The probability density functions of $F_\varepsilon(e)$ and $F_X(x)$ will be denoted, respectively, by $f_\varepsilon(e)$ and $f_X(x)$. The support of X is denoted by R_X and is supposed to be a compact subset of $\mathbb{R}$.

In order to estimate the distribution of ε, we first need to estimate $m(x)$ and $\sigma(x)$. For the regression function $m(x)$, following Cristóbal et al. (2004), we use a local linear estimator, adjusted for the selection bias in the following way:

$$\hat{m}_n(x)=\sum_{i=1}^n W_i^w(x,h_n)Y_i^w, \tag{4}$$

where

$$W_i^w(x,h_n)=\frac{w_i^w(x,h_n)}{\sum_{j=1}^n w_j^w(x,h_n)},$$

$$w_i^w(x,h_n)=\frac{1}{w_i}\left(s_2^w(x,h_n)K\left(\frac{X_i^w-x}{h_n}\right)-s_1^w(x,h_n)K\left(\frac{X_i^w-x}{h_n}\right)\frac{X_i^w-x}{h_n}\right),$$

$$s_k^w(x,h_n)=\sum_{i=1}^n\frac{1}{w_i}K\left(\frac{X_i^w-x}{h_n}\right)\left(\frac{X_i^w-x}{h_n}\right)^k$$