Statistics and Probability Letters 97 (2015) 88–98

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Robust inference in partially linear models with missing responses

Ana M. Bianco [a,b,*], Graciela Boente [c,d], Wenceslao González-Manteiga [e], Ana Pérez-González [f]

[a] Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, 1428, Buenos Aires, Argentina

[b] CONICET, Av. Rivadavia 1917, 1429, Buenos Aires, Argentina

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

[d] IMAS, CONICET, Ciudad Universitaria, Pabellón 1, 1428, Buenos Aires, Argentina

[e] Departamento de Estatística e Investigación Operativa, Facultad de Matemáticas, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain

[f] Departamento de Estadística e Investigación Operativa, Facultad de Empresariales y Turismo, Universidad de Vigo, Campus de Orense, 32004 Orense, Spain

ARTICLE INFO

Article history:
Received 13 March 2014
Received in revised form 22 October 2014
Accepted 7 November 2014
Available online 14 November 2014

MSC:
62G35
62F12
62F03

Keywords:
Kernel weights
Hypothesis testing
M-location functionals
Missing at random
Partly linear models
Robust estimation

ABSTRACT

We consider robust testing on the regression parameter of a partially linear regression model, where missing responses are allowed. We derive the asymptotic behavior of the proposed test statistic under the null and contiguous alternatives. A numerical study is performed.

1. Introduction

Non-parametric regression models suffer from the *curse of dimensionality* when the dimension of the covariates increases. Therefore, introducing some structure in the regression function the statistical analysis may become more efficient. Partially linear models (PLM) provide a solution to a large number of covariates by assuming that the regression function has two components: one depending linearly on some of the covariates, while the other one is non-parametric. In particular, PLM

* Correspondence to: Instituto de Cálculo, FCEyN, UBA, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina. Fax: +54 11 45763375.
E-mail addresses: abianco@dm.uba.ar (A.M. Bianco), gboente@dm.uba.ar (G. Boente), wenceslao.gonzalez@usc.es (W. González-Manteiga), anapg@uvigo.es (A. Pérez-González).

http://dx.doi.org/10.1016/j.spl.2014.11.004
0167-7152/

came to be more popular in the last years due to their flexibility, since the two components allow them to adapt to a wide class of situations. Sometimes, little is known about the relation among the response and some of the independent variables and hence, when the form of functional relation is unspecified, the use of a non-parametric component is recommended. In these situations, PLM are an appealing choice.

More formally, under a PLM, it is assumed that the response $y_i \in \mathbb{R}$ and the covariates $(\mathbf{x}_i^{\mathrm{T}}, t_i)$, $\mathbf{x}_i \in \mathbb{R}^p$, $t_i \in \mathbb{R}$, are such that

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + g(t_i) + \sigma\epsilon_i, \quad 1 \le i \le n, \tag{1}$$

where the errors ϵ_i are i.i.d., independent of $(\mathbf{x}_i^{\mathrm{T}}, t_i)$ with symmetric distribution $F_0(\cdot)$. That is, we assume that the error's scale equals 1 so as to identify the scale parameter as σ. We will not require any moment conditions on the errors distribution, but we only assume that the scale parameter for the errors equals 1. When the existence of seconds moments is assumed, as it is the case of the classical approach, these conditions imply that $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{VAR}(\epsilon_i) = 1$, which entails that, in this situation, σ represents the standard deviation of the responses conditional to the covariates.

Härdle et al. (2000, 2004) give an extensive description of different results obtained in PLM. In particular, in the context of hypothesis testing, Gao (1997) considers asymptotic test statistics for the problem $H_0 : \boldsymbol{\beta} = \mathbf{0}$, while González Manteiga and Aneiros Pérez (2003) studied the case of dependent errors. Classical procedures based on local polynomials and least squares estimation can be seriously damaged by a small fraction of anomalous observations. Robust estimates under the partly linear model were considered in He et al. (2002), where M-type estimates for repeated measurements using B-splines are introduced. On the other hand, Bhattacharya and Zhao (1997) define a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ by taking differences of the observations and combining a bandwidth-matched M-estimation procedure with kernel weights, when $p = 1$ and the carriers $\mathbf{x}$ lie in a compact set. Bianco and Boente (2004) introduce a kernel-based three-step procedure in order to achieve robustness against anomalous data including high leverage points in $\mathbf{x}$.

Nevertheless, in practice, not all the responses may be available, this may be planned or unplanned. The methods described above are designed for complete data sets and problems arise when missing observations are present. In some cases, people may refuse to provide some kind of information, in others, the response variable may be very expensive or difficult to measure. Also, sometimes there may be loss of information in the registration process or the researcher may fail to collect the full information. There are many situations in which both the response and the explanatory variables have missing values, however we will focus our attention on those cases where missing data occur only in the responses.

Wang et al. (2004) considered regression imputation of missing responses based on partly linear regression model in order to make inference on the mean of y. The estimator of $\boldsymbol{\beta}$, introduced by Wang et al. (2004), is a least squares regression estimator defined by considering preliminary kernel estimators, of the quantities $\mathbb{E}(\delta_1\mathbf{x}_1|t_1 = t)/\mathbb{E}(\delta_1|t_1 = t)$ and $\mathbb{E}(\delta_1 y_1|t_1 = t)/\mathbb{E}(\delta_1|t_1 = t)$, where $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ if y_i is missing. Estimators of the marginal mean of the response y based on the obtained estimator of the regression parameter are defined using an imputation estimator and also propensity score weighting estimators. Wang and Sun (2007) studied estimators of the regression coefficients and the nonparametric function using either imputation, semiparametric regression surrogate or an inverse marginal probability weighted approach. Since these estimators are based on weighted means of the response variables, they are highly sensitive to outliers. The lack of robustness of weighted means procedures pushed on the search of procedures resistant to outliers as those given in Bianco et al. (2010), who introduced robust estimators based on bounded score functions together with algorithms to compute them. In this paper, we go further and we focus our attention on inference regarding the parametric component, when the response variable has missing observations, but the covariates $(\mathbf{x}^{\mathrm{T}}, t)$ are totally observed.

The rest of the paper is organized as follows. Section 2 reviews the definition of the robust semiparametric estimators defined in Bianco et al. (2010) and recalls some previous results. In Section 3, the Wald test statistics are introduced, while their asymptotic distribution is derived under the null hypothesis and under contiguous alternatives in Section 3.1. The results of a simulation study are reported in Section 4, while some final comments are given in Section 5. Technical proofs are left to the Appendix.

2. Preliminaries

Consider a random sample of incomplete data $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i, \delta_i\right)$, $1 \le i \le n$, of a partially linear model where $\delta_i = 1$ if y_i is observed, $\delta_i = 0$ if y_i is missing, and the responses y_i satisfy model (1).

As mentioned above, our goal is to introduce robust tests to check hypotheses that engage the regression parameter $\boldsymbol{\beta}$ in the case where responses are possibly missing, in particular when they are missing at random (MAR). This means that if $(y, \mathbf{x}^{\mathrm{T}}, t, \delta)$ has the same distribution as $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i, \delta_i\right)$, δ is conditionally independent of the response y given $(\mathbf{x}^{\mathrm{T}}, t)$. In other words, we assume an ignorable mechanism such that $\mathbb{P}(\delta = 1|(y, \mathbf{x}^{\mathrm{T}}, t)) = \mathbb{P}(\delta = 1|(\mathbf{x}^{\mathrm{T}}, t)) = p(\mathbf{x}, t)$.

One may wonder if, ignoring the vectors with missing responses, we will still obtain robust and consistent procedures. That is, if the robust estimators given in Bianco and Boente (2004) applied to the observations $\{\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_N}\} = \{(y_i, \mathbf{x}_i^{\mathrm{T}}, t_i)^{\mathrm{T}}\}_{\delta_i=1}$, where $N = \sum_{i=1}^{n} \delta_i$, lead to asymptotically unbiased estimators so that, the tests defined through them in Bianco et al. (2006), turn out to be consistent. This is one of the conditions needed to successfully apply the transfer principle described in Koul et al. (2012). However, as mentioned in Bianco et al. (2010), a profile-likelihood procedure is needed to obtain consistent estimators for a wide class of situations when dealing with missing responses. Indeed, the robust estimators proposed in Bianco and Boente (2004) are not Fisher-consistent, unless the probability of missing responses is of the