Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Robust estimation in partially linear errors-in-variables models

Q2 Ana M. Bianco [a,*], Paula M. Spano [b]

[a] Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Buenos Aires, Argentina

[b] Departamento de Ciencias Exactas, Ciclo Básico Común, Universidad de Buenos Aires, Ciudad Universitaria, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form 31 August 2016
Accepted 1 September 2016
Available online xxxx

Keywords:
Fisher-consistency
Kernel weights
M-location functionals
Nonparametric regression
Robust estimation

ABSTRACT

In many applications of regression analysis, there are covariates that are measured with errors. A robust family of estimators of the parametric and nonparametric components of a structural partially linear errors-in-variables model is introduced. The proposed estimators are based on a three-step procedure where robust orthogonal regression estimators are combined with robust smoothing techniques. Under regularity conditions, it is proved that the resulting estimators are consistent. The robustness of the proposal is studied by means of the empirical influence function when the linear parameter is estimated using the orthogonal M-estimator. A simulation study allows to compare the behaviour of the robust estimators with their classical relatives and a real example data is analysed to illustrate the performance of the proposal.

1. Introduction

Two important branches of regression analysis arise from parametric and nonparametric models. The fully parametric models are readily interpretable, but they can be severely affected by misspecification. On the other hand, nonparametric models are very flexible to assess the relationship among variables, but they suffer from the well known *curse of dimensionality*. In the last decades semiparametric models, that amalgamate these two branches, have deserved a lot of attention. They take the best and avoid the worst of the parametric and nonparametric models. Among them, partially linear models have been extensively studied in the last years. Let $(y, \mathbf{x}^{\mathrm{T}}, t)$ be the observation in a subject or experimental unit, where y is the response that is related to the covariates $(\mathbf{x}^{\mathrm{T}}, t) \in \mathbb{R}^p \times \mathbb{R}$. The partially linear model assumes that

$$y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + g\,(t) + e,$$

where the error e is independent of the covariates $(\mathbf{x}^{\mathrm{T}}, t)$. By means of a nonparametric component, partially linear models are flexible enough to cover many situations; indeed, they can be a suitable choice when one suspects that the response y linearly depends on $\mathbf{x}$, but that it is nonlinearly related to t. An extensive description of the different results obtained in partially linear regression models can be found in Härdle et al. (2000). Among the robust literature, we find He et al. (2002) that consider M-type estimates for repeated measurements using B-splines and Bianco and Boente (2004) who introduce a kernel-based stepwise procedure to define robust estimates under a partially linear model.

[*] Correspondence to: Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Piso 2, 1428, Buenos Aires, Argentina. Fax: +54 11 45763375.

E-mail addresses: abianco@dm.uba.ar (A.M. Bianco), paumerspano@yahoo.com.ar (P.M. Spano).

http://dx.doi.org/10.1016/j.csda.2016.09.002
0167-9473/© 2016 Elsevier B.V. All rights reserved.

In practice, however, there often exist covariate measurement errors. This is a common situation in economics, medicine and social sciences. Errors-in-variables (EV) models have drawn a lot of attention and generated a wide literature, surveyed in Fuller (1999) and Carroll et al. (1995). The effect of measurement errors is well-known, indeed they can cause biased and inconsistent parameter estimators. Two approaches are adopted in order to overcome these difficulties according to the nature of the problem: the *functional* and *structural* modelling. In the functional model it is assumed that the covariates are deterministic, while in the *structural* model, which is treated in this paper, the covariates are considered as random variables. In our setting, we assume that we cannot observe $\mathbf{x}$ directly, but instead we observe a surrogate variable $\mathbf{v}$ which is related to $\mathbf{x}$ through the equation $\mathbf{v} = \mathbf{x} + \mathbf{e}_x$. In other words, the response and the vector of covariates $\mathbf{x}$ are observed with errors, while the scalar variable t is observable, that is, we assume the partially linear errors-in-variables (PLEV) model given by

$$\begin{aligned} y &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + g(t) + e, \\ \mathbf{v} &= \mathbf{x} + \mathbf{e}_x, \end{aligned} \tag{1}$$

where the vector of measurement errors

$$\boldsymbol{\epsilon} = \begin{pmatrix} e \\ \mathbf{e}_x \end{pmatrix} \tag{2}$$

is independent of $(\mathbf{x}^{\mathrm{T}}, t)$.

In order to correct for measurement error, some additional information or data is usually required. In the classical approach, at this point, there are two variants. In the first one, it is assumed that the covariance matrix of the measurement errors, $\boldsymbol{\Sigma}_{\mathbf{e}_x}$, is known and the approach is a correction for attenuation. Following these ideas, Liang et al. (1999) adapt the estimators of Severini and Staniswalis (1994), which combine local smoothers and linear parametric techniques, by including an attenuation term based on $\boldsymbol{\Sigma}_{\mathbf{e}_x}$ that enables to adjust the regression coefficients for the effects of measurement error. If $\boldsymbol{\Sigma}_{\mathbf{e}_x}$ were unknown, the estimation of the covariance matrix could be possible when replicates are available. In the second variant, it is assumed that the ratio between the variance of the error model e and the measurement errors $\mathbf{e}_x$ is known. This assumption allows for identification of the model. In this case, Liang et al. (1999) propose to estimate $\boldsymbol{\beta}$ by total least squares method.

Even when in practice the feasibility of any of these conditions depends on the problem, in general, in the robust framework assumptions involving the existence of first or second moments of the errors are avoided and replaced by weaker conditions on the errors distribution, such as symmetry. So, in this paper, we will extend the second variant by assuming that the vector of errors $\boldsymbol{\epsilon}$ follows a spherically symmetric distribution, which is a standard assumption in errors-in-variables models. In this case, if $\boldsymbol{\epsilon}$ has a density, it is of the form $\phi(\|\mathbf{u}\|)$ for some non-negative function ϕ. Spherical symmetry implies that e and each component of $\mathbf{e}_x$ have the same distribution. Cui and Kong (2006) justify this assumption by noticing that in some situations the response y and the covariate x are measured in the same way or, even more, the response and the non-observable covariate are two methods that measure the same quantity. As motivating example, we can consider the problem of predicting cholesterol serum level (CS) from a previous register of CS and age, which corresponds to the case of the real dataset we analyse below. First, it is sensible to assume that both cholesterol serum variables (the response and the covariate) are affected by an error, justifying to fit an EV model. Second, since both measures are of the same nature, it seems natural to assume that the errors of the response and the covariate follow the same distribution, making reasonable the sphericity assumption.

Among the literature in partially linear EV models, we can highlight the contribution of several authors. As mentioned, Liang et al. (1999) introduce a semiparametric version of the parametric correction for attenuation, while He and Liang (2000) consider consistent regression quantile estimates of $\boldsymbol{\beta}$. Partially linear models with measurement errors have been also studied by Ma and Carroll (2006), who propose locally efficient estimators in semiparametric models, Liang et al. (2007) that consider missing not at random responses, Pan et al. (2008) who deal with longitudinal data and by Liang and Li (2009) who focus on variable selection. As mentioned, we deal with the case in which variable t is observable. Measurement errors in both the parametric and the non-parametric part represent a much more complicated problem and would deserve a different approach, that is beyond the scope of this paper. In the classical setting, Liang (2000) and Zhu and Cui (2003), who deal with an unobservable variable t in the context of a partially linear model, consider deconvolution techniques to handle this type of situations.

However, if the smoothers involved in the estimation process are not resistant to outliers, then the resulting estimators can be severely affected by a relatively small fraction of atypical observations. The same can be asserted with respect to the estimation of the regression parameter when it is estimated by total least squares or least squares corrected for attenuation. For this reason, in this paper we consider an intuitively appealing way to obtain robust estimators for model (1) with spherically symmetric errors, which combines robust univariate smoothers with robust parametric estimators for a linear EV model. It is expected that the good robustness properties of estimates for linear EV models, such as M-orthogonal estimators or weighted orthogonal estimators introduced by Zamar (1989) and Fekri and Ruiz-Gazen (2004), respectively, combined with local smoothers, such as local medians or local M-type estimators, would result in estimators with good robustness properties as well. In what follows, we introduce a three-step procedure that yields robust and consistent estimators. We also derive the empirical influence function of the proposal when M-orthogonal estimators are used to estimate the regression parameter. The simulation results show that, regardless of the presence of outliers in the sample, the proposed estimators of the parametric and nonparametric components are very stable, making clear the advantage of using this kind of procedures.

Please cite this article in press as: Bianco, A.M., Spano, P.M., Robust estimation in partially linear errors-in-variables models. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.09.002