Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 1132 – 1142
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Local influence assessment in heteroscedastic measurement error models

Mário de Castro[a,*], Manuel Galea-Rojas[b], Heleno Bolfarine[c]

[a]*Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Caixa Postal 668, 13560-970, São Carlos-SP, Brazil*
[b]*Universidad de Valparaíso, Casilla 5030, Valparaíso, Chile*
[c]*Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281, Ag. Cidade de São Paulo, 05311-970, São Paulo-SP, Brazil*

Available online 17 May 2007

Abstract

Functional heteroscedastic measurement error models are investigated aiming to assess the effects of perturbations of data on some inferential procedures. This goal is accomplished by resorting to methods of local influence. The techniques provide to the practitioner a valuable tool that enables to identify potential influential elements and to quantify the effects of perturbations in these elements on results of interest. An illustrative example with a real data set is also reported.

Keywords: Regression analysis; Measurement error models; Maximum likelihood; Local influence

1. Introduction

Heteroscedastic measurement error models have received attention in the literature (see Fuller, 1987; Ripley and Thompson, 1987; Riu and Rius, 1996; Galea-Rojas et al., 2003; de Castro et al., 2004; Kukush and Van Huffel, 2004; de Castro et al., 2006a; Markovsky et al., 2006, among others). These works concentrate on parameter estimation and hypothesis testing. de Castro et al. (2006b) develop a local influence study, but they cover solely test statistics in a simple model (one response).

The chief aim of the present paper is the assessment of effects of minor perturbations of data on inferential results. To accomplish this goal we resort to methods of local influence. The roots of these methods are in the assessment of the individual impact of observations in a global sense, that is, an observation is either included or deleted in the analysis. There are three major points. First, meaningful perturbation schemes should be choosen in advance. Second, a selection of which particular aspects (for instance, parameter estimates) will be tracked under the perturbations. At last, an objective criterion to quantify the effects of perturbations.

The remaining of the paper is organized as follows. Besides the formulation of the working model, for the sake of completeness, maximum likelihood (ML) parameter estimation and hypothesis testing (as in de Castro et al., 2004) are sketched. Next, we provide a short account of the local influence assessment methodologies proposed by Cook (1986),

* Corresponding author. Tel.: +55 16 3373 9567; fax: +55 16 3373 9751.
E-mail address: mcastro@icmc.usp.br (M. de Castro).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2007.05.012

Wu and Luo (1993), and Cadigan and Farrell (2002), followed by the necessary matrices to implement some selected perturbation schemes. An illustrative example with a real data set is also reported.

2. Model and inference

Let n be the sample size; X_i, the unidimensional observed value of the covariate in unit i; Y_{ij}, the unidimensional jth observed response in unit i and x_i, the unobserved (true) covariate value for unit i. Relating these variables we postulate as working model the relations

$$X_i = x_i + u_i \tag{1}$$

and

$$Y_{ij} = \alpha_j + \beta_j x_i + e_{ij}, \tag{2}$$

$i = 1, \ldots, n$ and $j = 1, \ldots, r$. As an instance, this model is applicable to the comparison of measurement methods problem (Ripley and Thompson, 1987; Riu and Rius, 1996; Galea-Rojas et al., 2003; de Castro et al., 2004, among others). In this case, r is the number of methods to be compared to a reference one. Letting $\mathbf{e}_i = (e_{i1}, \ldots, e_{ir})^\top$, we assume that $\mathbf{e}_i$ is independent of u_i and are distributed as

$$(u_i, \mathbf{e}_i^\top)^\top \overset{\text{indep.}}{\sim} \mathrm{N}_{r+1}(\mathbf{0}, \mathbf{\Phi}_i), \tag{3}$$

$i = 1, \ldots, n$, where $\mathbf{\Phi}_i = \mathbf{D}(\kappa_i, \boldsymbol{\lambda}_i)$ denotes a diagonal matrix with elements κ_i and $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ir})^\top$. Expression (3) specifies the distribution of the $(r+1)$-dimensional vector of measurement errors. We suppose that the variances κ_i and $\boldsymbol{\lambda}_i$ are known and greater than 0, $i = 1, \ldots, n$, a common setup in areas such as Analytical Chemistry (Ripley and Thompson, 1987; Riu and Rius, 1996), Earth Science (Russell et al., 2000), Geochemistry (Maruyama and Yurimoto, 2003), and Water Science (Bertrand-Krajewski, 2004), to name just a few.

The model defined by Eqs. (1)–(2), can be written as

$$\mathbf{Z}_i = (X_i, \mathbf{Y}_{i\bullet}^\top)^\top = (0, \boldsymbol{\alpha}^\top)^\top + (1, \boldsymbol{\beta}^\top)^\top x_i + (u_i, \mathbf{e}_i^\top)^\top,$$

where $\mathbf{Y}_{i\bullet} = (Y_{i1}, \ldots, Y_{ir})^\top$, $i = 1, \ldots, n$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^\top$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)^\top$. Then, under assumption (3), it follows that

$$\mathbf{Z}_i \overset{\text{indep.}}{\sim} \mathrm{N}_{r+1}(\boldsymbol{\mu}_i, \mathbf{\Phi}_i),$$

where $\boldsymbol{\mu}_i = (x_i, \boldsymbol{\alpha}^\top + x_i\boldsymbol{\beta}^\top)^\top$, $i = 1, \ldots, n$.

The above model is known as a functional errors-in-variables model (Fuller, 1987; Cheng and Van Ness, 1999). Because no assumption is made on the distribution of the unknown values x_i, $i = 1, \ldots, n$, they are also parameters which have to be estimated. Since their number increases with the sample size, they are known as incidental parameters. As our main interest is on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $x_1, \ldots, x_n$ are nuisance parameters. Model (1)–(2) is a particular instance of the multivariate element-wise weighted total least squares (EW-TLS) (Kukush and Van Huffel, 2004; Markovsky et al., 2006).

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$, which is of dimension $2r \times 1$. The log-likelihood function corresponding to the model defined by (1) and (2) with assumption (3) can be written as

$$l(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}, \boldsymbol{x}), \tag{4}$$

where

$$l_i(\boldsymbol{\theta}, \boldsymbol{x}) = -\frac{r+1}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{\Phi}_i) - \frac{1}{2}Q_i(\boldsymbol{\theta}),$$