# Influence diagnostics in linear and nonlinear mixed-effects models with censored data

Larissa A. Matos [a], Victor H. Lachos [a,*], N. Balakrishnan [b], Filidor V. Labra [a]

[a] *Department of Statistics, Campinas State University, Campinas, São Paulo, Brazil*

[b] *Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada*

## ARTICLE INFO

## ABSTRACT

HIV RNA viral load measures are often subjected to some upper and lower detection limits depending on the quantification assays, and consequently the responses are either left or right censored. Linear and nonlinear mixed-effects models, with modifications to accommodate censoring (LMEC and NLMEC), are routinely used to analyze this type of data. Recently, Vaida and Liu (2009) proposed an exact EM-type algorithm for LMEC/NLMEC, called the SAGE algorithm (Meng and Van Dyk, 1997), that uses closed-form expressions at the E-step, as opposed to Monte Carlo simulations. Motivated by this algorithm, we propose here an exact ECM algorithm (Meng and Rubin, 1993) for LMEC/NLMEC, which enables us to develop local influence analysis for mixed-effects models on the basis of conditional expectation of the complete-data log-likelihood function. This is because the observed data log-likelihood function associated with the proposed model is somewhat complex which makes it difficult to directly apply the approach of Cook (1977, 1986). Some useful perturbation schemes are also discussed. Finally, the results obtained from the analyses of two HIV AIDS studies on viral loads are presented to illustrate the newly developed methodology.

## 1. Introduction

Studies of HIV viral dynamics, often considered to be a key issue in AIDS research, considers repeated/longitudinal measures over a period of treatment routinely analyzed using linear and non-linear mixed-effects models (LME/NLME) to assess rates of changes in HIV-1 RNA level or viral load (Wu, 2005, 2010). Viral load measures the amount of actively replicating virus and its reduction is frequently used as a primary endpoint in clinical trials of anti-retroviral (ARV) therapy. However, depending on the diagnostic assays used, its measurement may be subjected to some upper and lower detection limits, below or above which they are not quantifiable (resulting in left or right censoring). The proportion of censored data in these studies may not be small (Hughes, 1999) and so the use of crude/ad hoc methods, such as substituting a threshold value or some arbitrary point like a mid-point between zero and cut-off for detection (Vaida and Liu, 2009), might lead to biased estimates of fixed effects and variance components (Wu, 2010).

Our motivating datasets in this study are on HIV-1 viral loads, (i) after unstructured treatment interruption, or UTI (Saitoh et al., 2008) and (ii) setpoint for acutely infected subjects from the AIEDRP program (Vaida and Liu, 2009). The former has about 7% of observations below (left-censored) the detection limits, whereas the latter has about 22% lying above (right-censored) the limits of assay quantifications. As an alternative to crude imputation methods, Hughes (1999) proposed a

---

* Correspondence to: Departamento de Estatística, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, São Paulo, Brazil.
  *E-mail address:* hlachos@ime.unicamp.br (V.H. Lachos).

likelihood-based Monte Carlo EM algorithm (MCEM) for LME with censored responses (LMEC). Vaida et al. (2007) proposed a hybrid EM using a more efficient Hughes's algorithm, extending it to NLME with censored data (NLMEC). Recently, Vaida and Liu (2009) proposed an exact EM-type algorithm for LMEC/NLMEC, which uses closed-form expressions at the E-step, as opposed to Monte Carlo simulations. Strictly speaking, these algorithms are Space Alternating Generalized EM (SAGE) algorithms (see Vaida et al., 2007). In this paper, for performing diagnostics analysis in LMEC/NLMEC models, we first propose a slight modification to the EM-type algorithm proposed by Vaida and Liu (2009), wherein all the parameters are updated (M-step) by considering the random effects and the censored observations as missing data. Then, the diagnostic measures for assessing the local influence in LMEC/NLMEC are developed and presented.

The study of influence analysis is an important and key step in data analysis subsequent to parameter estimation. This can be carried out by conducting an influence analysis for detecting influential observations. There are two primary approaches for detecting influential observations. The first approach is the case-deletion approach (Cook, 1977) and it is an intuitively appealing method (see also Cook and Weisberg, 1982). Deletion diagnostics such as Cook's distance or the likelihood distance have been applied to many statistical models. The second approach, which is a general statistical technique used to assess the stability of the estimation outputs with respect to the model inputs, is the local influence approach of Cook (1986). Following the pioneering work of Cook (1986), this method has received considerable attention recently in the statistical literature on mixed-effects models (LME/NLME); see, for example, Lesaffre and Verbeke (1998), Zhu and Lee (2001), Lee and Xu (2004), Osorio et al. (2007) and Russo et al. (2009), among others.

Although several diagnostic studies on LME/NLME have appeared in the literature, to the best of our knowledge, no study seems to have been made on influence diagnostics for NLMEC/LMEC and certainly not on the local influence analysis. The main difficulty is due to the fact that the observed log-likelihood function of the NLMEC/LMEC involves intractable integrals (for instance, the pdfs of truncated multinormal distributions), rendering the direct application of Cook's approach (Cook, 1986) to this model to be very difficult if not impossible, since the measures involve the first and second partial derivatives of this function. Zhu and Lee (2001) developed an approach for performing local influence analysis for general statistical models with missing data, and it is based on the $Q$-displacement function that is closely related to the conditional expectation of the complete-data log-likelihood in the E-step of the EM algorithm. This approach produces results very similar to those obtained from Cook's method. Moreover, the case-deletion can be studied by $Q$-displacement function following the approach of Zhu et al. (2001). So, we develop here methods to obtain case-deletion measures and local influence measures by using the method of Zhu et al. (2001) (see also Lee and Xu, 2004; Zhu and Lee, 2001) in the context of mixed-effects models with censored data. It is our opinion that the results developed here form a necessary supplement to those presented by Vaida and Liu (2009) for the analysis of mixed-effects models with censored response and HIV data.

The rest of this paper is organized as follows. In Section 2, the LMEC model is defined, and an EM-type algorithm for obtaining the ML estimates is described. In Section 3, we provide a brief sketch of the local influence approach for models with incomplete data, and also develop a methodology pertinent to the LMEC. Four different perturbation schemes are considered. In Section 4, the NLMEC model is defined. The methodology has been illustrated in Section 5 with the analysis of two examples involving HIV viral measure and by an empirical study. Finally, some concluding remarks are made in Section 6.

## 2. The linear mixed-effects with censored response

Ignoring censoring for the moment, the classical normal LME model is specified as follows (Laird and Ware, 1982):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{b}_i \overset{\text{i.i.d.}}{\sim} N_q(\mathbf{0}, \mathbf{D})$ is independent of $\boldsymbol{\epsilon}_i \overset{\text{ind.}}{\sim} N_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i})$, $i = 1, \ldots, n$; the subscript $i$ is the subject index; $\mathbf{I}_p$ denotes the $p \times p$ identity matrix; $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is an $n_i \times 1$ vector of observed continuous responses for subject $i$; $\mathbf{X}_i$ is the $n_i \times p$ design matrix corresponding to the fixed effects, $\boldsymbol{\beta}$, of dimension $p \times 1$; $\mathbf{Z}_i$ is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects $\mathbf{b}_i$; $\boldsymbol{\epsilon}_i$ of dimension ($n_i \times 1$) is the vector of random errors; and the dispersion matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ depends on unknown and reduced parameters $\boldsymbol{\alpha}$. In the present formulation, we consider the case where the response $Y_{ij}$ is not fully observed for all $i, j$. Let the observed data for the $i$-th subject be $(\mathbf{V}_i, \mathbf{C}_i)$, where $\mathbf{V}_i$ represents the vector of uncensored reading or censoring level, and $\mathbf{C}_i$ is the vector of censoring indicators such that

$$\begin{aligned} y_{ij} &\leq V_{ij} \quad \text{if } C_{ij} = 1, \\ y_{ij} &= V_{ij} \quad \text{if } C_{ij} = 0. \end{aligned} \tag{2}$$

For simplicity, we will assume that the data are left-censored with the LMEC being thus defined. The extensions of theses results to arbitrary censoring can be easily presented.

### 2.1. The log-likelihood function

Following Vaida and Liu (2009), classical inference on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$ is based on the marginal distribution of $\mathbf{y}_i$. For complete data, we have marginally that $\mathbf{y}_i \overset{\text{ind.}}{\sim} N_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$. For responses