Diagnostic and treatment for linear mixed models

Julio M. Singer[1,4], **Juvêncio S. Nobre**[2] and **Francisco M. M. Rocha**[3]

[1] Universidade de São Paulo, São Paulo, BRAZIL
[2] Universidade Federal do Ceará, Fortaleza, BRAZIL
[3] Universidade Federal de São Paulo, São José dos Campos, BRAZIL
[4] Corresponding author: Julio M Singer, e-mail: jmsinger@ime.usp.br

Abstract

We consider residual, local influence and leverage analyses to identify violations of the assumptions underlying gaussian linear mixed models. In particular, we propose different diagnostic measures to analyze marginal, conditional and random-effects residuals and develop R-based software to implement such tools. We propose remedial measures that range from fine-tuning of the model to the adoption of more robust elliptically symmetric distributions as well as generalized linear mixed models or GEE-based models and comment on the available diagnostic tools for such alternatives. We consider practical examples where some of the underlying assumptions are invalid, show how such violations are detected with the proposed tools and redefine the models to accommodate them. Finally, we suggest directions for further research
Keywords: local influence, global influence, residual analysis, remedial measures.

1. Introduction

The linear mixed model (LMM) may be expressed as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \ldots, n \tag{1}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$ is a $m_i \times 1$ vector of observations (**response profile**) for the i-th unit, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of unknown population parameters (**fixed effects**), $\mathbf{X}_i$ is a $m_i \times p$ known specification matrix corresponding to the fixed effects, $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^\top$ is a $q \times 1$ vector of unobservable random elements (**random effects**), $\mathbf{Z}_i$ is a $m_i \times q$ known specification matrix corresponding to the random effects and $\mathbf{e}_i = (e_{i1}, \ldots, e_{im_i})^\top$ is an $m_i \times 1$ vector of random errors. We assume that the $\mathbf{b}_i$ and the $\mathbf{e}_i$ are uncorrelated and that $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$, $\mathbb{V}(\mathbf{b}_i) = \mathbf{G}$, $\mathbb{E}(\mathbf{e}_i) = \mathbf{0}$, $\mathbb{V}(\mathbf{e}_i) = \mathbf{R}_i$, where $\mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$ and $\mathbf{R}_i = \mathbf{R}_i(\boldsymbol{\theta})$ are, respectively, $q \times q$ and $m_i \times m_i$ positive-definite symmetric matrices depending on an $r \times 1$ covariance parameter vector $\boldsymbol{\theta}$, not functionally related to $\boldsymbol{\beta}$. Under this setup, $\mathbf{y}_i$ has mean vector $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $\mathbb{V}(\mathbf{y}_i) = \boldsymbol{\Omega}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^\top + \mathbf{R}_i$. Letting $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top)^\top$, $\mathbf{Z} = \oplus_{i=1}^n \mathbf{Z}_i$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$ and $\mathbf{e} = (\mathbf{e}_1^\top, \ldots, \mathbf{e}_n^\top)^\top$, we can write model (1) more compactly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \tag{2}$$

and therefore, $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbb{V}(\mathbf{y}) = \boldsymbol{\Omega} = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}^\top + \boldsymbol{\Sigma}$ where $\boldsymbol{\Gamma} = \mathbf{I}_n \otimes \mathbf{G}$ and $\boldsymbol{\Sigma} = \oplus_{i=1}^n \mathbf{R}_i$.

Assuming that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are known, best linear unbiased estimators (BLUE) of $\boldsymbol{\beta}$ and best linear predictors (BLUP) of $\mathbf{b}_i$ may be obtained as the solutions to the Henderson equations, namely

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y} \text{ and } \widehat{\mathbf{b}} = \boldsymbol{\Gamma}\mathbf{Z}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{\Gamma}\mathbf{Z}^\top\mathbf{Q}\mathbf{y}, \tag{3}$$

where $\mathbf{Q} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}$. Then $\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, $\mathbb{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$ and $\mathbb{E}(\widehat{\mathbf{b}}) = \mathbf{0}$, $\mathbb{V}(\widehat{\mathbf{b}}) = \boldsymbol{\Gamma}\mathbf{Z}^\top\mathbf{Q}\mathbf{Z}\boldsymbol{\Gamma}$. In practice, empirical BLUE or BLUP may be obtained by replacing $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ with consistent estimates in (3). It is common to assume that both $\mathbf{b}_i$ and $\mathbf{e}_i$ in (1) follow independent gaussian distributions,

If $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ were known, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ would be

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{X}^\top[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\mathbf{X})^{-1}\mathbf{X}^\top[\boldsymbol{\Omega}(\boldsymbol{\theta})]^{-1}\mathbf{y}. \tag{4}$$

To reduce the bias in the estimation of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$, the Restricted Maximum Likelihood Estimator (REMLE), $\widehat{\boldsymbol{\theta}}_R$ is considered and substituted for $\boldsymbol{\theta}$ in (4). Large sample theory methods may be used to show that $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}_R) \approx N_p\{\boldsymbol{\beta}, (\mathbf{X}^\top[\boldsymbol{\Omega}(\widehat{\boldsymbol{\theta}_R})]^{-1}\mathbf{X})^{-1}\}$. The reader is referred to Demidenko (2004), among others.

2. Diagnostic in the gaussian setup

Residual and sensitivity analyses constitute important tools for evaluating the fit of any statistical model to given data, for checking the validity of its assumptions and consequently, to evaluating the reliability of statistical inference based on it.

Residual analysis is more complex in LMM because the associated sources of variation ($\mathbf{e}$ and $\mathbf{b}$) generate the following three types of residuals

i) **Marginal residuals**, $\widehat{\boldsymbol{\xi}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, that predict the marginal errors $\boldsymbol{\xi} = \mathbf{y} - \mathbb{E}[\mathbf{y}] = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$.

ii) **Conditional residuals**, $\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}\widehat{\mathbf{b}}$, that predict the conditional errors $\mathbf{e} = \mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{b}] = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}$.

iii) **BLUPs**, $\mathbf{Z}\widehat{\mathbf{b}}$, that predict the random effects, $\mathbf{Z}\mathbf{b} = \mathbb{E}[\mathbf{y}|\mathbf{b}] - \mathbb{E}[\mathbf{y}]$.

Hilden-Minton (1995) considers a residual confounded for a specific type of error if it depends on other errors than the one that it is supposed to predict. In particular, conditional residuals and the BLUP may be confounded and so that $\widehat{\mathbf{e}}$ may not be adequate to check for the normality of $\mathbf{e}$ because when $\mathbf{b}$ is grossly non-gaussian, $\widehat{\mathbf{e}}$ may not present a gaussian behaviour even when $\mathbf{e}$ is gaussian.

Lesaffre and Verbeke (1998) comment that when the within-unit covariance structure is adequate, $\mathcal{V}_i = ||\mathbf{I}_{n_i} - \widehat{\boldsymbol{\mathcal{R}}}_i\widehat{\boldsymbol{\mathcal{R}}}_i^\top||^2$, where $\widehat{\boldsymbol{\mathcal{R}}}_i = \widehat{\boldsymbol{\Omega}}_i^{-1/2}\widehat{\boldsymbol{\xi}}_i$ with $\widehat{\boldsymbol{\Omega}}_i = \boldsymbol{\Omega}_i(\widehat{\boldsymbol{\theta}})$ should be close to zero. We suggest a slight modification of their proposal, obtained by replacing $\widehat{\boldsymbol{\mathcal{R}}}_i$ in $\mathcal{V}_i$ with the standardized marginal residuals $\widehat{\boldsymbol{\xi}}_i^* = [\widehat{\mathbb{V}}(\widehat{\boldsymbol{\xi}}_i)]^{-1/2}\widehat{\boldsymbol{\xi}}_i$ where $\widehat{\mathbb{V}}(\widehat{\boldsymbol{\xi}}_i)$ corresponds to the diagonal block of $\widehat{\boldsymbol{\Omega}} - \mathbf{X}(\mathbf{X}^\top\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top$ associated to the i-unit and using the standardized index $\mathcal{V}_i^* = \mathcal{V}_i/\sum_{i=1}^n \mathcal{V}_i$ instead of $\mathcal{V}_i$, to facilitate comparison between different models.

Given that $\mathbb{V}(\widehat{\mathbf{e}}) = \boldsymbol{\Sigma}[\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}]\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{Q}\boldsymbol{\Sigma}$, we also modify the proposal of Nobre and Singer (2007) and consider standardized conditional residuals defined as as $\widehat{\mathbf{e}}_i^* = [\widehat{\mathbb{V}}(\widehat{\mathbf{e}}_i)]^{-1/2}\widehat{\mathbf{e}}_i$ where $\widehat{\mathbb{V}}(\widehat{\mathbf{e}}_i)$ corresponds to the diagonal block of $\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{Q}}\widehat{\boldsymbol{\Sigma}}$ associated to the i-th unit. Hilden-Minton (1995) advocates the use of **least confounded conditional residuals**, *i.e.*, of a linear transformation of the conditional residuals that minimizes the fraction of confounding.

When there is no confounding and the random effects follow a q-dimensional gaussian distribution, $\mathcal{M}_i = \widehat{\mathbf{b}}_i^\top\{\widehat{\mathbb{V}}[\widehat{\mathbf{b}}_i - \mathbf{b}_i]\}^{-1}\widehat{\mathbf{b}}_i$ (the Mahalanobis's distance between $\widehat{\mathbf{b}}_i$ and $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$) should have a chi-squared distribution with q degrees of freedom.

The different uses for the three types of LMM residuals is summarized in Table 1, adapted from Nobre and Singer (2007).

In models like (2) a unit or a (within-unit) observation can affect both marginal and conditional fitted values; therefore it seems reasonable to evaluate the joint influence of each unit or (within-unit) observation on both. In this context, the **generalized joint leverage matrix** is

$$\mathbf{L} = \partial\widehat{\mathbf{y}}^*/\partial\mathbf{y}^\top = \partial\widehat{\mathbf{y}}/\partial\mathbf{y}^\top + \partial\mathbf{Z}\widehat{\mathbf{b}}/\partial\mathbf{y}^\top = \mathbf{L}_1 + \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}^\top\mathbf{Q} = \mathbf{L}_1 + \mathbf{L}_2\mathbf{Q}, \tag{5}$$

where $\widehat{\mathbf{y}}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$, $\mathbf{L}_1 = \mathbf{X}\left(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}$ and $\mathbf{L}_2 = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}^\top$. Nobre and Singer (2011) observe that $\mathbf{H}_2 = \mathbf{Z}\boldsymbol{\Gamma}\mathbf{Z}^\top\mathbf{Q} = \mathbf{L}_2\boldsymbol{\Omega}^{-1}\left[\mathbf{I}_n - \mathbf{L}_1\right]$ also depends on $\mathbf{L}_1$, and argue that the leverage of the conditional fitted values with respect to the random effects may be affected by the leverage with respect to the marginal fitted values and suggest using $\mathbf{L}_2$ instead of $\mathbf{H}_2$ to measure leverage with respect to the random effects. The diagonal blocks, $\mathbf{L}_{1i} = \mathbf{X}_i(\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}_i^\top\boldsymbol{\Omega}_i^{-1}$ and $\mathbf{L}_{2i} = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^\top$ of $\mathbf{L}_1$ and $\mathbf{L}_2$ as well

Table 1: Uses of residuals for diagnostic purposes

Diagnostic for	Type of residual	Plot
Linearity of effects fixed ($\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$)	Marginal	$\widehat{\boldsymbol{\xi}}_i^*$ *vs* fitted values or explanatory variables
Presence of outlying observations	Marginal	$\widehat{\boldsymbol{\xi}}_i^*$ *vs* observation indices
Within-subjects covariance matrix ($\boldsymbol{\Omega}_i$)	Marginal	$\mathcal{V}_i^*$ *vs* unit indices
Presence of outlying observations	Conditional	$\widehat{e}_k^*$ *vs* observation indices
Homoskedasticity of conditional errors ($\mathbf{e}_i$)	Conditional	$\widehat{e}_k^*$ *vs* fitted values
Normality of conditional errors ($\mathbf{e}_i$)	Conditional	Gaussian QQ plot for $\widehat{\mathbf{e}}_k^*$ or $\mathbf{c}_k^\top \widehat{\mathbf{e}}^*$
Presence of outlying subjects	EBLUP	$\mathcal{M}_i^*$ *vs* unit indices
Normality of the random effects ($\mathbf{b}_i$)	EBLUP	χ^2_q QQ plot for $\mathcal{M}_i$

as their diagonal elements $\mathbf{L}_{1i(jj)}$ and $\mathbf{L}_{2i(jj)}$ may be employed to evaluate the leverage of units or (within-unit) observations with respect to either the fixed or random terms of the model.

In the spirit of **case deletion**, to examine the impact of the set of units $I = \{i_1, i_2, ..., i_k\}$ $(1 \leq i_1 \leq i_2 \leq ... \leq i_k \leq n)$, on some characteristic of interest in linear mixed models, we may consider the augmented model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{U}_I\boldsymbol{\phi}_{(I)} + \mathbf{e}, \tag{6}$$

where $\boldsymbol{\phi}_I$ represents a k-dimensional (fixed) parameter vector and $\mathbf{U}_I = [\mathbf{u}_{i_1}, \mathbf{u}_{i_2}, ..., \mathbf{u}_{i_k}]$ with $\mathbf{u}_i$ representing the i-th column of $\mathbf{I}_N$. When $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are known, Fung et al. (2002) show that the BLUE of $\boldsymbol{\beta}$ and the BLUP of $\mathbf{b}$ in model (6) or in model (2) with deletion of the units in I are equal and denote them by $\widehat{\boldsymbol{\beta}}_{(I)}$ and $\widehat{\mathbf{b}}_{(I)}$, respectively. They also show that $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)} = (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{U}_I\widehat{\boldsymbol{\phi}}_{(I)}$ where $\widehat{\boldsymbol{\phi}}_{(I)} = (\mathbf{U}_I^\top\mathbf{Q}\mathbf{U}_I)^{-1}\mathbf{U}_I^\top\mathbf{Q}\mathbf{y}$, obtain the variance of $\mathbb{V}(\widehat{\boldsymbol{\beta}}_{(I)})$, define Cook's distance $D_I = (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)})^\top\boldsymbol{\Omega}^{-1}(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(I)})/p$ and use it as a measure of influence of the units in the set I on the estimate of $\boldsymbol{\beta}$.

Tan et al. (2001), among others, propose a conditional approach based on observation-oriented influence measures. They assume that the covariance matrix of $\mathbf{e}_i$ in (1) is $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$, consider the conditional model $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{e}$ where $\mathbf{X}^* = [\mathbf{X}\ \mathbf{Z}]$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^\top, \mathbf{b}^\top)^\top$ and define the **conditional Cook distance** as

$$D_{i(j)}^{cond} = \textstyle\sum_{i=1}^n(\widehat{\mathbf{y}}_i^* - \widehat{\mathbf{y}}_{i(j)}^*)^\top(\widehat{\mathbf{y}}_i^* - \widehat{\mathbf{y}}_{i(j)}^*)/[\sigma^2(n+p)]. \tag{7}$$

where $\widehat{\mathbf{y}}_i^* = \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}$, $\widehat{\mathbf{y}}_{i(j)}^* = \mathbf{X}_i\widehat{\boldsymbol{\beta}}_{(i(j))} + \mathbf{Z}_i\widehat{\mathbf{b}}_{(i(j))}$ and $\widehat{\boldsymbol{\beta}}_{(i(j))}$ and $\widehat{\mathbf{b}}_{(i(j))}$ denote, respectively the BLUEs of $\boldsymbol{\beta}$ and $\mathbf{b}$ obtained with the elimination of the j-th observation from the i-th unit. The $D_{i(j)}^{cond}$ may be decomposed in three terms, $D_{1i(j)}^{cond}$, $D_{2i(j)}^{cond}$ and $D_{3i(j)}^{cond}$, observation index plots of which are useful for evaluating the impact of the deletion of the j-th observation from the i-th unit on different components of the model.

The ratio of the variance ellipsoids, $\rho_{(I)} = |\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}_{(I)})|/|\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}})|$ can also be used to evaluate the influence of the units in I on the covariance matrix of $\widehat{\boldsymbol{\beta}}$ as highlighted by Hilden-Minton (1995).

Global influence may be evaluated by the index-plots summarized in Table 2.

Local influence is used to investigate the behaviour of the **likelihood displacement**, $LD(\boldsymbol{\omega}) = 2\{L(\widehat{\boldsymbol{\psi}}) - L(\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}|\boldsymbol{\omega})\}$ where L denotes the likelihood for the proposed model, $\boldsymbol{\psi}$ is a p-dimensional parameter vector, $\boldsymbol{\omega} \in \Omega \subset I\!R^q$ is a q-dimensional vector of small "perturbations" and $\widehat{\boldsymbol{\psi}}$ and $\widehat{\boldsymbol{\psi}}_{\boldsymbol{\omega}}$ are, respectively, the MLE of $\boldsymbol{\psi}$ based on $L(\boldsymbol{\psi})$ or on $L(\boldsymbol{\psi}|\boldsymbol{\omega})$. Such perturbations may be imposed on the response variable,

Table 2: Global influence plots for observations or units

Diagnostic for effect on	Global influence measure	Index-plot of
Fixed portion of fitted value ($\mathbf{X}\widehat{\boldsymbol{\beta}}$)	Generalized marginal leverage matrix $\mathbf{L}_1$	$\mathbf{L}_{1i(jj)}$ $[tr(\mathbf{L}_{1i})/m_i]$ *vs* observations (units)
Random portion of fitted value ($\mathbf{Z}\widehat{\mathbf{b}}$)	Generalized random component marginal leverage matrix $\mathbf{L}_2$	$\mathbf{L}_{2i(jj)}$ $[tr(\mathbf{L}_{2i})/m_i]$ *vs* observations (units)
Regression coefficients ($\widehat{\boldsymbol{\beta}}$)	Cook's distance D_I	$D_{i(j)}$ $[D_i]$ *vs* observations (units)
Covariance matrix of regression coefficients $[\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}})]$	Ratio of variance ellipsoids $\rho_{(I)}$	$\rho_{i(j)}$ $[\rho_i]$ *vs* observations (units)

i.e., $\mathbf{y}_i(\boldsymbol{\omega}_i) = \mathbf{y}_i + \boldsymbol{\omega}_i$, on the explanatory variables, *i.e.*, $\mathbf{X}_i(\mathbf{W}_i) = \mathbf{X}_i + \mathbf{W}_i$, where $\mathbf{W}_i = [\boldsymbol{\omega}_{i1}, \ldots, \boldsymbol{\omega}_{ip}]$ with $\boldsymbol{\omega}_{ij} = (\omega_{ij1}, \ldots, \boldsymbol{\omega}_{ijm_i})^\top$, on the variance of the random effects, *i.e.*, $\mathbf{G}(\omega_i) = \omega_i\mathbf{G}$ or on the variance of the errors, $\mathbf{R}_i(\omega_i) = \omega_i\mathbf{R}_i$. As the selection of the appropriate perturbation scheme in LMM is not straightforward, Lesaffre and Verbeke (1998) adopt a slightly different and more practical approach for local influence diagnostics using the marginal likelihood of $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, namely $L(\boldsymbol{\psi}) = \sum_{i=1}^n L_i(\boldsymbol{\psi})$ and including perturbations of its individual terms by taking $L_i(\boldsymbol{\psi}_{\boldsymbol{\omega}}) = \sum_{i=1}^n \omega_i L_i(\boldsymbol{\psi})$ where L_i denotes the likelihood for the i-th unit and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$.

3. Treatment

Diagnostic tools are useful to highlight the inadequacy of some characteristics of tentative models but they rely on the correct specification of the within-unit covariance structure as well as on the underlying distributions. We consider some remedial approaches when these tools suggest that the proposed models do not accommodate one or more of the aspects of the data.

To refine the within-unit covariance structure in the case of random polynomial coefficient models, Rocha and Singer (2012) suggest i) simple t-tests based on the estimated coefficients of standard linear regression models fitted to each unit's data as a tool for selecting fixed effects and ii) Bonferroni-corrected reference intervals for selecting random effects. They also note that that when $\mathbf{R}_i = \sigma^2\mathbf{I}_{m_i}$, the structure of the columns of the within-unit covariance matrix is similar to that of the individual profiles. When the data for all subjects are collected at the same time points, $\mathbf{V} = \mathbf{Z}^*\mathbf{G}\mathbf{Z}^{*\top}$ may be easily estimated and hey suggest using the same approach adopted for the individual profiles on the rows of $\widehat{\mathbf{V}}$ to decide which random effects should be included in the model. Based on a simulation study, they show that the procedure is reasonably efficient even for moderate sample sizes.

Grady and Helms (1995) suggest that plots of covariances and correlations *versus* time between measurements (lags) may be used as a tool for identifying possible autoregressive covariance patterns and propose a strategy to compare different models.

Pinheiro et al. (2001), Savalli et al. (2006), Osorio et al. (2007), Bolfarine et al. (2007), among others, consider **elliptically-symmetric** (ES) or **skew-elliptical** (SE) distributions to bypass the lack of robustness of standard LMM. Residual and leverage analyses are still not well established for such models. However, the similarity with the gaussian case suggests some exploratory tools. For example, index-plots of the weighted marginal residuals may be used to detect outliers as suggested by Savalli et al. (2006). For the most common non-gaussian ES distributions, the weights tend to reduce the influence of units associated to larger values of Mahalanobis's distance and therefore may accommodate outliers. A similar analysis may be carried out for the conditional residuals. Local influence is considered by Osorio et al. (2007).

The effects of asymmetry on the appropriateness of gaussian theory methods are, in

general, more serious than those of heavy tails and SE distributions may be considered as alternatives. In practice it is quite complicated to fit models in this class with the exception of skew-normal (SN) distributions. Bolfarine et al. (2007) consider the LMM (1) with underlying SN distributions and propose local influence diagnostic measures. They also consider a case-deletion diagnostic measure and mention that evaluating model adequacy is still an open problem.

Generalized linear mixed models (GLMM) allow non-gaussian response distributions as well as a possible non-linear relations between the expected response and explanatory variables. In this context, Xiang et al. (2002) develop diagnostics based on Cook's distance to evaluate the impact of deleting an unit on the MLE of the fixed parameters. Their approach is extended to accommodate joint and conditional influence measures, but is suitable to handle at most two influential units. Xu et al. (2006) mention that robust procedures could be considered to address the effects of influential units or observations, but that this, as well as the corresponding diagnostic tools are still unavailable for this class of mixed models.

GEE-based models, on the other hand, focus directly on the marginal distribution of the data $\mathbf{y}$, with no reference to random effects. Such models may not be classified as mixed models, but they play an important role in the analysis of repeated measures or longitudinal data. Venezuela et al. (2007) define residuals in this setup and consider a "hat" matrix similar to the one employed in standard linear models to detect high-leverage units and obtain the equivalent to Cook's distance. They propose an algorithm to generate half-normal plots with simulated envelopes that are useful to identify outliers and evaluate the adequacy of the model.

More recently, Venezuela et al. (2011) propose local influence diagnostics for a larger class of regression models that include those for which the **fit function** (*e.g*, likelihood function or quasi-likelihood function) $\mathcal{F}(\boldsymbol{\beta})$ is not specified. Under this setup, they investigate the behaviour of the **fit function displacement**, $2[\mathcal{F}(\widehat{\boldsymbol{\beta}}) - \mathcal{F}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}})]$ where $\boldsymbol{\omega}$ is a perturbation vector.

4. Computation and examples

The numerous approaches to the analysis of repeated measures via LMM may turn out to be a source of concern when fitting such models to real data. First, it is difficult to decide which class of models and what analysis strategy to adopt. Second, suitable software is not always available or may not accommodate the peculiarities that accompany practical problems. The modelling strategy should include an iterative procedure according to which after each new model is fitted, appropriate diagnostic tools should be employed to check whether the new proposal is more adequate than the previous one. The most popular statistical software packages for fitting LMM are `SAS proc MIXED` and the libraries `lme4` and `nlme` in the `R` software package. `R`-functions for diagnostics in gaussian LMM as well as some examples may be obtained from www.ime.usp.br/∼jmsinger/LMMdiagnostics.zip.

5. Discussion

Gaussian LMM are very flexible, easily interpretable and may be fitted via a series of very efficient algorithms. If both the fixed and random components are well specified, the results obtained in practical applications are usually very similar to those generated by other classes of models as highlighted in Pinheiro et al. (2001), Savalli et al. (2006) or Alencar et al. (2012), for example.

Using the diagnostic tools in this setup may not be an easy task because many functions designed to generate diagnostic plots are still not implemented in the most commonly used statistical software packages. Although some of these functions may be obtained from the authors, their use in practical applications may not be a straightforward task. In particular, we mention extracting the necessary information from the fitting function output, specially for more complex models. This is even more problematic for the other classes of models (ES LMM and SE LMM, GLMM and GEE-based

models), where even the definition of residuals is not well established. Some effort in designing and testing flexible functions for such purposes could help to identify cases where the use of the gaussian LMM may not be adequate and to disseminate the use of models that can accommodate outlying or influential observations.

An additional difficulty relates to interpretation of the results and to their incorporation into new versions of the model. It is not clear how to compare the results of different classes of models. The standard errors of the fixed parameters tend to be smaller when models based on heavy-tailed distribution are fitted, but we have no idea of possible biases. Good simulation studies should be conducted to clarify these issues.

Acknowledgements

This research received financial support from CNPq and FAPESP, Brazil.

References

Alencar, A.P., Singer, J.M. and Rocha, F.M.M. (2012). Competing regression models for longitudinal data. *Biometrical Journal*, **54**, 214-229.

Bolfarine, H., Montenegro, L.C. and Lachos, V.H. (2007). Influence diagnostics for skew-normal linear mixed models. *Sankhya*, **69**, 648-670.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: John Wiley & Sons.

Fung, W.K., Zhu, Z.Y., Wei, B.C. and He, X. (2002). Influence diagnostics and outliers tests for semiparametric mixed models. *Journal of the Royal Statistical Society, B*, **64**, 565–579.

Grady, J.J. and Helms, R.W. (1995). Model selection techniques for the covariance matrix for incomplete longitudinal data. *Statistics in Medicine*, **14**, 1397–1416.

Hilden-Minton, J.A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models*. Unpublished PhD Thesis. University of California, Los Angeles.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Nobre, J.S. and Singer, J.M. (2007). Residual analysis for linear mixed models. *Biometrical Journal*, **49**, 863–875.

Nobre, J.S. and Singer, J.M. (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics*, **38**, 1063–1072.

Osorio, F., Paula, G.A. and Galea, M. (2007). Assessment of local influence in elliptical linear models with longitudinal structure. *Computational Statistics and Data Analysis*, **51**, 4354-4368.

Pinheiro, J.C., Liu, C. and Wu, Y.N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. *Journal of Computational and Graphical Statistics*, **10**, 249–276.

Rocha, F.M.M. and Singer, J.M. (2012). Selection of terms in random coefficients models. *Submitted*.

Savalli, C., Paula, G.A. and Cysneiros, F.J.A. (2006). Assessment of variance components in elliptical linear mixed models. *Statistical Modelling*, **6**, 59–76.

Tan, F.E.S., Ouwens, M.J.N. and Berger, M.P.F. (2001). Detection of influential observations in longitudinal mixed effects regression models. *The Statistician*, **50**, 271–284.

Venezuela, M.K., Botter, D.A. and Sandoval, M.C. (2007). Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation*, **77**, 879-888.

Venezuela, M.K., Sandoval, M.C. and Botter, D.A. (2011). Local influence in estimating equations. *Computational Statistics and Data Analysis*, **55**, 1867-1883.

Xiang, L., Tse, S-K. and Lee, A.H. (2002). Influence diagnostics for generalized linear mixed models: applications to clustered data. *Computational Statistics and Data Analysis*, **40**, 759-774.

Xu, L., Lee, S-Y and Poon, W-Y. (2006). Deletion measures for generalized linear mixed effects models. *Computational Statistics and Data Analysis*, **51**, 1131-1146.