

## Competing regression models for longitudinal data

Airlane P. Alencar<sup>\*1</sup>, Julio M. Singer<sup>1</sup>, and Francisco Marcelo M. Rocha<sup>2</sup>

<sup>1</sup> Departamento de Estatística, Instituto de Matemática e Estatística, Universidade of São Paulo, Caixa Postal 66281, São Paulo, SP05314-970, Brazil

<sup>2</sup> Departamento de Ciência e Tecnologia, Universidade Federal de São Paulo, São José dos Campos, SP, Brazil

Received 04 March 2011; revised 24 August 2011; accepted 16 December 2011

The choice of an appropriate family of linear models for the analysis of longitudinal data is often a matter of concern for practitioners. To attenuate such difficulties, we discuss some issues that emerge when analyzing this type of data via a practical example involving pretest–posttest longitudinal data. In particular, we consider log-normal linear mixed models (LNLMM), generalized linear mixed models (GLMM), and models based on generalized estimating equations (GEE). We show how some special features of the data, like a nonconstant coefficient of variation, may be handled in the three approaches and evaluate their performance with respect to the magnitude of standard errors of interpretable and comparable parameters. We also show how different diagnostic tools may be employed to identify outliers and comment on available software. We conclude by noting that the results are similar, but that GEE-based models may be preferable when the goal is to compare the marginal expected responses.

*Keywords:* Estimating equations method; Generalized linear models; Longitudinal data; Mixed models; Pretest/posttest measures.

Supporting Information for this article is available from the author or on the WWW under <http://dx.doi.org/10.1002/bimj.201100056>

### 1 Introduction

Mixed models are very useful to analyze data with a hierarchical structure, where measures from distinct units are independent and those from within units are correlated. In particular, they are appropriate to analyze repeated measures in longitudinal studies. The classical linear mixed model considered by Laird and Ware (1982) is expressed as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \dots, N, \quad (1)$$

where  $\mathbf{y}_i$  is the  $(n_i \times 1)$  vector of responses for the  $i$ -th unit,  $\mathbf{X}_i$  is the  $(n_i \times p)$  fixed effects specification matrix,  $\boldsymbol{\beta}$  is the corresponding  $(p \times 1)$  vector of parameters,  $\mathbf{b}_i$  is a  $(q \times 1)$  vector of random effects with  $E(\mathbf{b}_i) = \mathbf{0}$  and  $\text{Var}(\mathbf{b}_i) = \mathbf{G}$ ,  $\mathbf{Z}_i$  is a  $(n_i \times q)$  specification matrix for the random effects, and  $\mathbf{e}_i$  is a  $(n_i \times 1)$  vector of random errors independent of  $\mathbf{b}_i$ , with  $E(\mathbf{e}_i) = \mathbf{0}$ ,  $\text{Var}(\mathbf{e}_i) = \mathbf{R}_i$ , and  $\mathbf{G}$  ( $q \times q$ ) and  $\mathbf{R}_i$  ( $n_i \times n_i$ ) are symmetric positive definite matrices. The marginal expected response is  $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$  and the conditional (on the individual effect  $\mathbf{b}_i$ ) expected response is  $E(\mathbf{y}_i|\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ . The inclusion of random effects may account for heterogeneity of regression coefficients across sample units, possibly due to several unmeasured factors that affect the response variable (Fitzmaurice et al., 2004).

<sup>\*</sup>Corresponding author: e-mail: lane@ime.usp.br, Phone: +55-11-30916218, Fax: 55 11 30916130

It follows that

$$\Sigma = \text{Var}(\mathbf{y}_i) = \mathbf{Z}_i \mathbf{G} \mathbf{Z}_i^\top + \mathbf{R}_i, \quad (2)$$

where the first and second summands correspond, respectively, to the inter- and intraindividual covariance structures. When  $\mathbf{R}_i = \sigma^2 \mathbf{I}$ , the model is termed homoskedastic conditionally independent model, and correlation among observations on the same unit  $i$  arises from their sharing the unobservable (latent) variable,  $\mathbf{b}_i$  (Diggle et al., 2002, p. 128).

In the original formulation, Laird and Ware (1982) assumed that  $\mathbf{b}_i$  and  $\mathbf{e}_i$  in (1) follow normal distributions. Unbiased and consistent estimators of the variance components may be obtained by maximizing the restricted likelihood function, and estimators of the fixed effects may be obtained via maximum likelihood (ML) (Jiang, 2007). In practice, however, it is often not plausible to assume that the response is normally distributed. To bypass this problem, a possible strategy is to consider a log-normal linear mixed model (LNLMM), where the response is assumed to follow a multiplicative model with log-normal errors; the corresponding linearized model is additive and may be expressed as (1). For a broader choice of response distributions, two alternatives are generalized linear mixed models (GLMM) and models in which the parameters are estimated via generalized estimating equations (GEE).

In the GLMM approach, a distribution of the exponential family is considered for the response. In the usual formulation, individual observations are assumed to be independent conditionally on the random effects. In a more general setting, it is possible to consider different covariance structures. Estimation may be based on ML, penalized quasi-likelihood (PQL), or pseudo-likelihood methods (see Fitzmaurice et al. [2008, pp. 90–94] for details).

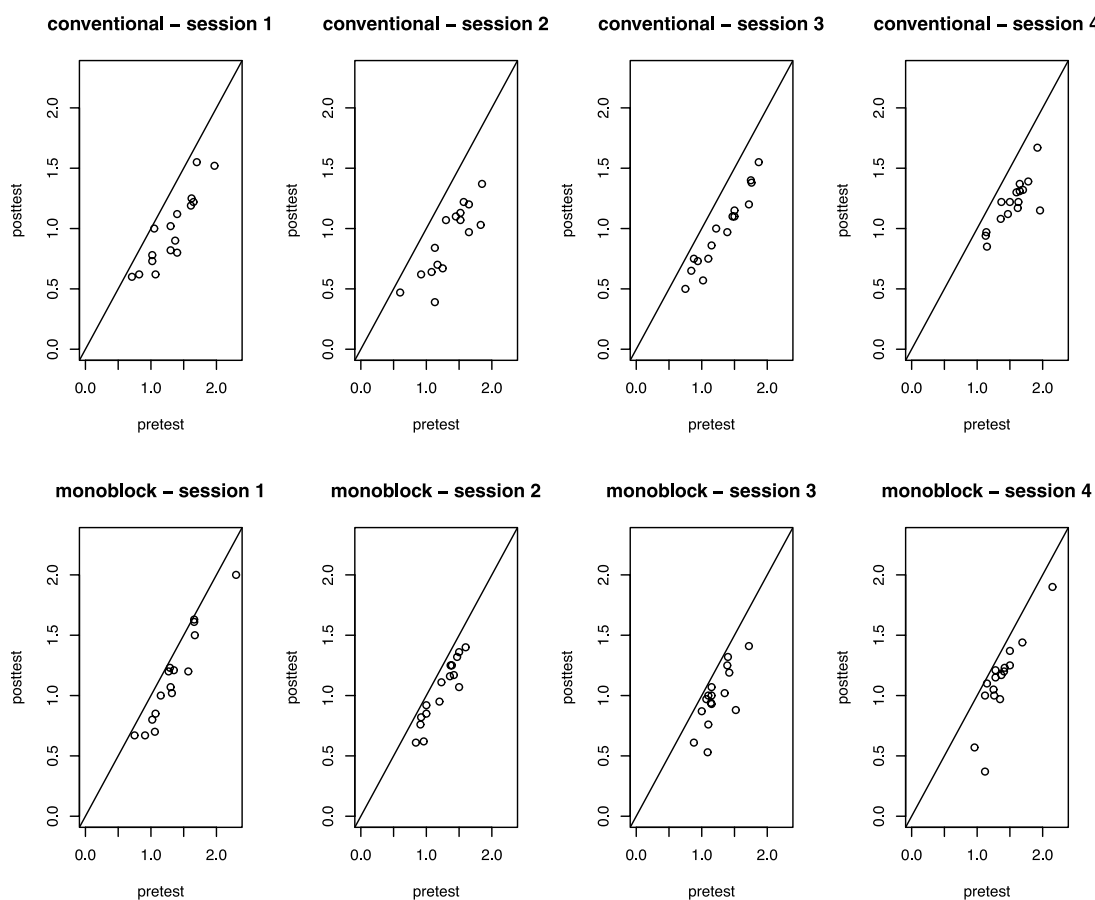
In the GEE-based approach, the form of the distribution of the vectors of responses ( $\mathbf{y}_i$ ) is not specified; the only assumption is that the marginal model depends only on the mean vector and on the covariance matrix of  $\mathbf{y}_i$ . Estimators are obtained as solutions to GEE and are consistent even when the covariance structure is misspecified. In this case, the standard errors may not be correct, but valid standard errors can be obtained via the sandwich estimator (Liang and Zeger, 1986). Because GEE-based models do not belong to the class of mixed models, they do not allow the evaluation of individual effects.

Excellent surveys on linear and GLMM are presented in Demidenko (2004) and Jiang (2007). Marginal models analyzed via GEE are carefully addressed in Hardin and Hilbe (2002). Song (2007) and Fitzmaurice et al. (2008) present recent reviews on the all these models. Good references on longitudinal data analysis using mixed and marginal models are Diggle et al. (2002), Fitzmaurice et al. (2004) and Molenberghs and Verbeke (2005).

We compare some features of the three approaches by fitting LNLMM-, GLMM-, and GEE-based models to data obtained from a pretest–posttest longitudinal study. We show how some important characteristics of the data may be incorporated in the three models, discuss the robustness of parameter estimators in the presence of outliers, and comment on their computational implementation. In Section 2, we introduce the data along with a preliminary exploratory analysis with the objective of identifying models for the analysis. In Sections 3, 4, and 5, we present the results of fitting, respectively, LNLMM-, GLMM-, and GEE-based models, detailing their specification, the estimation method, and discussing diagnostic tools. All codes are available in the supplemental material. In Section 6, we discuss some advantages, disadvantages, and properties of each family of models and present some concluding remarks.

## 2 Preliminary analysis of the longitudinal pretest–posttest data

We present a pretest–posttest longitudinal data set, highlighting its distinctive characteristics and conduct a preliminary descriptive analysis. In particular, we evaluate coefficients of variation and fit a naive model using least squares in order to identify possible competing models.



**Figure 1** Scatter plots of pretest versus posttest bacterial plaque indices.

A study conducted at the School of Dentistry of the University of São Paulo, Brazil, was designed to compare the efficiency in the removal of bacterial plaque under daily use of a low cost experimental (monoblock) toothbrush to that of a conventional toothbrush. In the study, 32 children, four to six years old, were randomly divided into two groups. Sixteen children received the monoblock toothbrush and the remaining children received the conventional toothbrush at the beginning of the study and used it for 45 days. During this period, data on a bacterial plaque index were collected before (pretest) and after (posttest) toothbrushing in sessions spaced by 15 days. The data is available in Nobre and Singer (2007).

Scatter plots of the pretest and posttest bacterial plaque indices for each session and type of toothbrush are presented in Fig. 1.

The bacterial plaque indices measured before ( $x$ ) and after ( $y$ ) toothbrushing are supposed to have the following characteristics, suggested in Singer and Andrade (1997) which we include for the sake of self-containment:

- (a) A pretest plaque index equal to zero implies an expected posttest plaque index also equal to zero.
- (b) Pretest and posttest plaque indices are non-negative.

- (c) The data are possibly heteroskedastic (because the response is non-negative and follows the relation  $E(y) \leq x$ ); i.e., the expected posttest bacterial plaque index ( $y$ ) must be smaller than the pretest index ( $x$ ).
- (d) The relation between pretest and posttest plaque indices may be non-linear.
- (e) The observations carried out on the same child are possibly correlated.

For simplicity, we analyze the data conditionally on the values of the pretest observations. For an error-in-variables approach to a similar problem, the reader is referred to Aoki et al. (2003).

Following the suggestions of Singer and Andrade (1997) and Singer et al. (2002), a model that satisfies (a)–(e) is a multiplicative model of the form

$$y_{ijd} = \beta_{jd} x_{ijd}^{\gamma_{jd}} \varepsilon_{ijd}, \quad (3)$$

where  $y_{ijd}$  is the posttest bacterial plaque index for the  $i$ -th individual using the  $j$ -th toothbrush in the  $d$ -th session,  $x_{ijd}$  is the corresponding pretest index,  $i = 1, \dots, 16$ ,  $j = 0$  (conventional), 1 (monoblock),  $d = 1, 2, 3, 4$ ,  $\beta_{jd}$  and  $\gamma_{jd}$  are parameters to be estimated and  $\varepsilon_{ijd}$  are non-negative random errors.

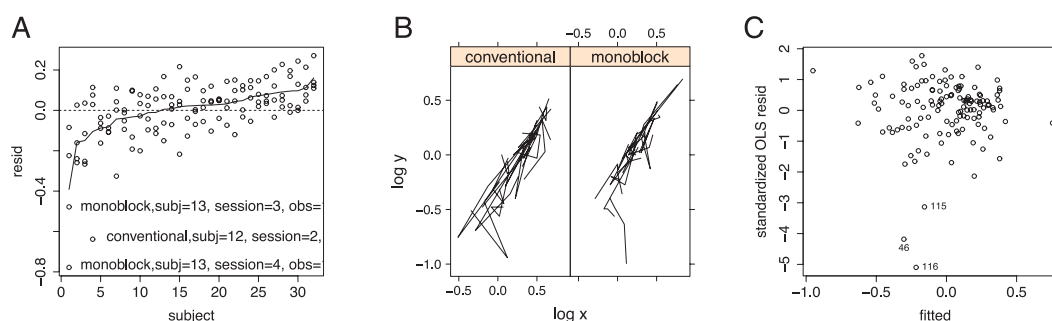
The efficiency of a toothbrush may be measured by the relative residual bacterial plaque index, defined as the expected ratio between the posttest and pretest bacterial plaque indices, say  $E(y_{ijd}|x_{ijd}) = \beta_{jd} x_{ijd}^{\gamma_{jd}-1} E(\varepsilon_{ijd})$ . With this definition, the smaller the relative residual bacterial plaque index, the more efficient is the toothbrush. Note that the measure of efficiency is proportional to  $\beta_{jd}$  and it is equal to  $\beta_{jd}$  when  $\gamma_{jd} = 1$  and  $E(\varepsilon_{ijd}) = 1$ . In this multiplicative model, if  $\gamma_{jd} = 1$ , the efficiency does not depend on the pretest index, but when  $\gamma_{jd} > 1$  ( $\gamma_{jd} < 1$ ), the efficiency decreases (increases) according to the pretest index.

The characteristics (a)–(d) are clearly satisfied by model (3); in particular the heteroskedasticity for the response stems from  $\text{Var}(y_{ijd}) = (\beta_{jd} x_{ijd}^{\gamma_{jd}})^2 \text{Var}(\varepsilon_{ijd})$ , even for homoskedastic errors ( $\varepsilon_{ijd}$ ). Requirement (e) may be incorporated in the specification of the error distribution. Model (3) may be linearized by taking logarithms, i.e.,

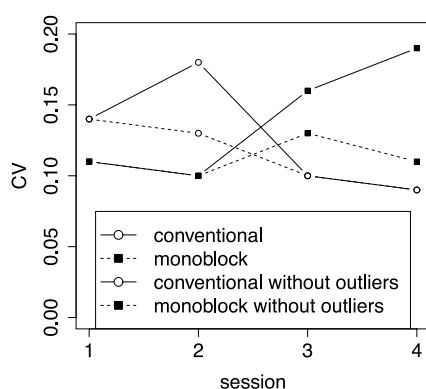
$$\log(y_{ijd}) = \lambda_{jd} + \gamma_{jd} \log(x_{ijd}) + e_{ijd}, \quad (4)$$

where  $\lambda_{jd} = \log(\beta_{jd})$  and  $e_{ijd} = \log(\varepsilon_{ijd})$ .

To identify whether random intercepts should be included, we started by fitting model (4) assuming uncorrelated errors. In Fig. 2A, we plot the residuals versus the rank of the mean residual (mean of the four residuals for each child); some children have all four residuals larger than others, suggesting



**Figure 2** (A) Residuals versus ranks of subject mean residuals. (B) Individual profiles of  $\log(y)$  versus  $\log(x)$ . (C) Residuals versus fitted values.



**Figure 3** Sample coefficients of variation of ratios between the posttest and pretest indices for each toothbrush and session for the complete and reduced (excluding three outliers) data sets.

that there are subject-specific components affecting the response. This variability may be modeled by random intercepts. The individual profiles of  $\log(y)$  versus  $\log(x)$  plotted in Fig. 2B do not suggest that individual slopes are different so that random slopes may not be necessary. Standardized residuals are plotted versus fitted values in Fig. 2C. Three outliers may be identified; they correspond to the second evaluation for the 12-th child in the conventional toothbrush group and for the third and fourth evaluation sessions of the 13-th child in the monoblock toothbrush group.

The analysis of the parameter estimates of model (4) also suggests that  $\gamma_{jd} = 1$ . Under this restriction, the efficiency of a toothbrush in each session can be measured by the expected posttest/pretest bacterial plaque index ratio,  $E(y_{ijd}/x_{ijd}) = \beta_{jd}E(\varepsilon_{ijd})$ . Since the corresponding standard deviation is  $SD(y_{ijd}/x_{ijd}) = \beta_{jd}SD(\varepsilon_{ijd})$ , the coefficient of variation  $CV(y_{ijd}/x_{ijd}) = SD(\varepsilon_{ijd})/E(\varepsilon_{ijd})$  should be constant for all sessions and toothbrush groups if the errors  $\varepsilon_{ijd}$  were identically distributed. Figure 3 suggests that the coefficients of variation of the ratios  $y_{ijd}/x_{ijd}$  are different, and consequently that the standard deviations of the errors  $\varepsilon_{ijd}$  in (3) may be larger for the conventional toothbrush in sessions 1 and 2 and for the monoblock toothbrush in sessions 3 and 4. Even after removing the three outliers suggested in Fig. 2C, the coefficients of variation still exhibit a similar behavior.

The preliminary analysis suggests that the candidate models should include random intercepts, heteroskedasticity, and a larger coefficient of variation for the ratio posttest/pretest bacterial plaque indices for the conventional toothbrush in sessions 1 and 2 and for the monoblock toothbrush in sessions 3 and 4.

### 3 Analysis via LNLMM

Here we consider a LNLMM model of the form (4) that includes random effects to account for possible positive within-subjects correlations and also considers an error covariance matrix that incorporates the nonconstant coefficients of variation identified in Section 2. Such models may be fitted via well established linear mixed model methodology, where the location and the covariance structure may be modeled separately.

The LNLMM may be expressed as

$$y_{ij}^* = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{1}_4 b_{ij} + \mathbf{e}_{ij}, \quad i = 1, \dots, 16, j = 0, 1, \quad (5)$$

**Table 1** Expressions for the estimators of  $\beta_{jd}^* = E(y_{ijd}|x_{ijd} = 1)$  and corresponding asymptotic variances.

Model	$\widehat{\beta}_{jd}^*$	$\text{Var}(\widehat{\beta}_{jd}^*)$
LNLMM	$\widehat{\beta}_{jd} \exp[0.5(\widehat{\sigma}_b^2 + \widehat{r}_{jd}^2)]$	$(\beta_{jd}^*)^2 [\text{Var}(\widehat{\lambda}_{jd}) + 0.25(\sigma_b^2 + r_{jd}^2)]$
GLMM	$\widehat{\beta}_{jd} \exp(0.5\widehat{\sigma}_b^2)$	$(\beta_{jd}^*)^2 [\text{Var}(\widehat{\lambda}_{jd}) + 0.25(\sigma_b^2)]$
GEE	$\widehat{\beta}_{jd}$	$(\beta_{jd}^*)^2 [\text{Var}(\widehat{\lambda}_{jd})]$

where  $\mathbf{y}_{ij}^* = (\log(y_{ij1}), \log(y_{ij2}), \log(y_{ij3}), \log(y_{ij4}))^\top$ , the specification matrices for the conventional and monoblock toothbrush groups ( $j = 0, 1$ ) are respectively,

$$\mathbf{X}_{i0} = \left( \mathbf{I}_4, \mathbf{0}_4, \bigoplus_{d=1}^4 \log(x_{ijd}), \mathbf{0}_4 \right), \mathbf{X}_{i1} = \left( \mathbf{0}_4, \mathbf{I}_4, \mathbf{0}_4, \bigoplus_{d=1}^4 \log(x_{ijd}) \right),$$

where  $\bigoplus_{i=1}^4 a_i$  denotes a diagonal matrix with the elements  $a_i$  along the main diagonal,  $\boldsymbol{\beta} = [\boldsymbol{\lambda}^\top \boldsymbol{\gamma}^\top]^\top$  is a  $(16 \times 1)$  vector with  $\boldsymbol{\lambda}^\top = [\lambda_{01}, \lambda_{02}, \lambda_{03}, \lambda_{04}, \lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14}]$ ,  $\boldsymbol{\gamma}^\top = [\gamma_{01}, \gamma_{02}, \gamma_{03}, \gamma_{04}, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}]$ ,  $\mathbf{I}_4 = (1, 1, 1, 1)^\top$ ,  $\mathbf{I}_4$  denotes the identity matrix of order 4, and  $\mathbf{0}_4$  denotes a  $4 \times 4$  matrix with all elements equal to zero. The error vectors  $\mathbf{e}_{ij}$  are independent and follow  $N(\mathbf{0}, \mathbf{R}_j)$  distributions, where, as suggested by the preliminary analysis,

$$\mathbf{R}_j = \bigoplus_{d=1}^4 r_{jd}^2 = \begin{cases} \text{diag}\{\tau_1^2, \tau_1^2, \tau_2^2, \tau_2^2\}, & \text{if } j = 0 \\ \text{diag}\{\tau_2^2, \tau_2^2, \tau_1^2, \tau_1^2\}, & \text{if } j = 1 \end{cases} \tag{6}$$

Also, we assume that the random effects,  $b_{ij}$ , are independent and follow  $N(0, \sigma_b^2)$  distributions.

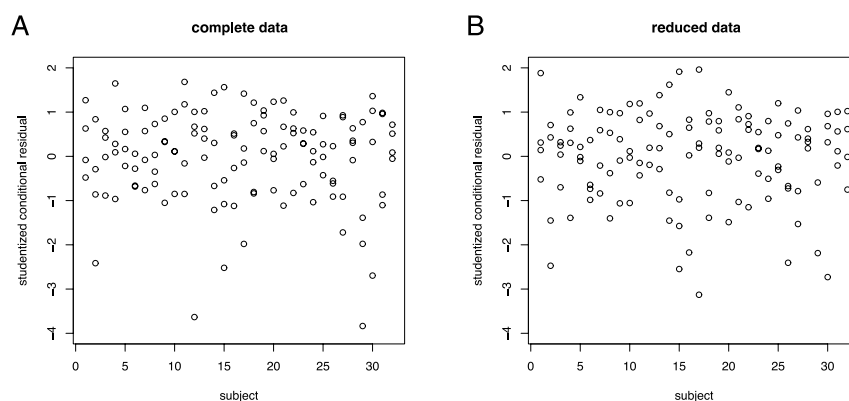
We present estimates of  $\beta_{jd}^* = E(y_{ijd}|x_{ijd} = 1)$  instead of the estimates of  $\beta_{jd} = \exp(\lambda_{jd})$  in (3) as suggested in Singer et al. (2002). To compute these estimates, note that  $\exp(b_{ij} + e_{ijd})$  has a log-normal distribution, so that  $E[\exp(b_{ij} + e) + e_{ijd}] = \exp[(\sigma_b^2 + r_{jd}^2)/2]$ . This approach is more convenient for interpretation and comparison with other models. Expressions for estimators  $\widehat{\beta}_{jd}^*$  and their corresponding asymptotic variances (obtained by the delta method) are presented in Table 1. Model (5) and (6) was fitted to the data using restricted ML methods, available in SAS PROC MIXED (Littell et al., 2006). The codes and the data to fit the proposed models may be provided online in the link of Supporting Information. The estimates are presented in Table 2.

Results obtained under this model were compared to those based on two other models of the form (5): the first with  $\mathbf{R}_j = \text{diag}(\tau_{j1}^2, \dots, \tau_{jd}^2)$ ,  $j = 0, 1$ , and the second with  $\mathbf{R}_j = \tau^2 \mathbf{I}_4$ . This latter was the model adopted in Nobre and Singer (2007) in their analysis and implies a constant coefficient of variation. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) obtained via restricted ML for the three models were respectively (AIC = -96.8 and BIC = -92.4), (AIC = -91.8, BIC = -78.6), and (AIC = -74.9 and BIC = -72.0), suggesting that (5)–(6) is acceptable. The reader is referred to Jiang and Rao (2003) for a detailed analysis of generalized information criteria and to Guerin and Stroup (2000) or Gurka (2006) for a discussion on AIC and BIC for model comparison.

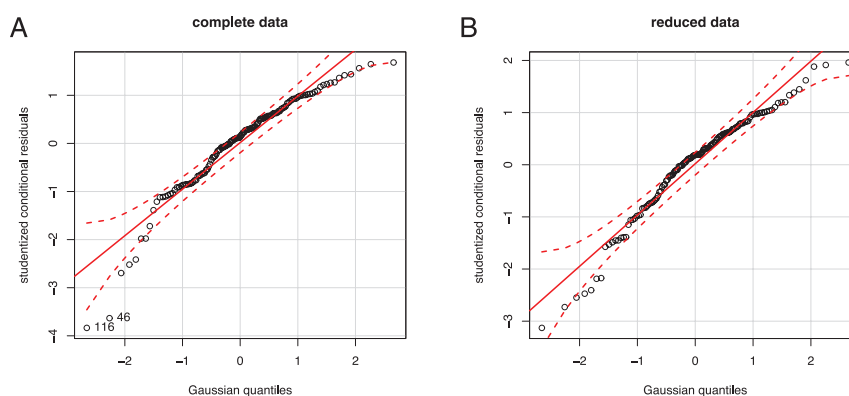
In order to evaluate the robustness of the estimators, the model was refitted to the reduced data obtained by omitting the three outliers suggested in Fig. 2C. The results are also presented in Table 2. The major discrepancies between the estimates for the complete and reduced data occur for the parameters  $\beta_{jd}^*$  and  $\gamma_{jd}$  for  $(j, d) = (0, 2), (1, 3), (1, 4)$ , which correspond to the treatments

**Table 2** Estimates (Est.) and standard errors (SE) for saturated LNLMM, GLMM, and GEE fitted to the complete and reduced data sets.

Parameter	LNLMM				GLMM				GEE			
	Complete data		Reduced data		Complete data		Reduced data		Complete data		Reduced data	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
$\beta_{01}^*$	0.768	0.052	0.757	0.034	0.758	0.043	0.755	0.032	0.769	0.032	0.759	0.030
$\beta_{02}^*$	0.666	0.047	0.705	0.034	0.669	0.040	0.703	0.032	0.671	0.041	0.704	0.024
$\beta_{03}^*$	0.738	0.024	0.736	0.024	0.737	0.024	0.735	0.023	0.744	0.027	0.738	0.026
$\beta_{04}^*$	0.872	0.059	0.872	0.058	0.866	0.056	0.866	0.056	0.856	0.043	0.861	0.044
$\beta_{11}^*$	0.820	0.029	0.817	0.029	0.819	0.028	0.815	0.028	0.826	0.030	0.817	0.032
$\beta_{12}^*$	0.824	0.029	0.825	0.028	0.822	0.028	0.822	0.027	0.836	0.029	0.828	0.030
$\beta_{13}^*$	0.805	0.061	0.835	0.044	0.797	0.052	0.831	0.042	0.807	0.047	0.834	0.039
$\beta_{14}^*$	0.709	0.067	0.801	0.054	0.725	0.059	0.802	0.051	0.748	0.084	0.810	0.067
$\gamma_{01}$	0.941	0.178	0.954	0.120	0.938	0.155	0.952	0.115	0.892	0.097	0.946	0.082
$\gamma_{02}$	1.014	0.174	0.932	0.118	0.991	0.152	0.933	0.113	0.985	0.120	0.935	0.073
$\gamma_{03}$	1.050	0.088	1.061	0.088	1.045	0.085	1.056	0.085	1.013	0.082	1.048	0.077
$\gamma_{04}$	0.759	0.148	0.758	0.147	0.769	0.142	0.769	0.142	0.799	0.144	0.785	0.144
$\gamma_{11}$	1.144	0.095	1.160	0.094	1.141	0.091	1.159	0.091	1.115	0.084	1.155	0.096
$\gamma_{12}$	1.110	0.120	1.105	0.120	1.103	0.116	1.103	0.116	1.024	0.106	1.080	0.111
$\gamma_{13}$	1.003	0.291	0.901	0.198	1.008	0.254	0.918	0.190	0.963	0.211	0.907	0.204
$\gamma_{14}$	1.450	0.270	1.173	0.188	1.367	0.236	1.153	0.180	1.282	0.240	1.127	0.193
$\sigma^2$	0.004	0.002	0.003	0.002	0.004	0.002	0.003	0.002				
$\tau_1^2$	0.035	0.007	0.007	0.002	0.026	0.006	0.014	0.003				
$\tau_2^2$	0.007	0.002	0.015	0.003	0.007	0.002	0.007	0.002				
$\alpha$									0.323	0.118	0.258	0.088
$\phi$									0.009	0.002	0.008	0.002
$\phi_1$									0.019	0.009	0.007	0.003



**Figure 4** Studentized conditional residuals versus subject indices for LNMM. (A) Complete data. (B) Reduced data.



**Figure 5** QQ plot of studentized conditional residuals for LNMM. (A) Complete data. (B) Reduced data.

with identified outliers. As expected, the estimated error variance ( $\widehat{\sigma}_b^2$ ) is smaller for the reduced data set.

Following the suggestions of Nobre and Singer (2007), studentized conditional residual plots are presented in Fig. 4 for both the complete and reduced data; the two outlying observations in Fig. 4A correspond to large decreases in the posttest index relatively to the pretest index in the second session for the 12-th child in the conventional toothbrush group and in the fourth session for the 13-th child in the monoblock toothbrush group. It is worth noting that fitting the model with  $\mathbf{R}_j = \tau^2 \mathbf{I}_4$ , as in Nobre and Singer (2007), these two outliers are more apparent in the residual analysis; furthermore, an extra outlier is detected in the third session for a child in the monoblock group (as in Fig. 2C). The model that includes (6) accommodates this observation and seems to be more suitable. The QQ plots in Fig. 5A (complete data) and 5b (omitting possible outliers) suggest mild deviations from the normality assumption, but we do not believe that this could jeopardize the results.

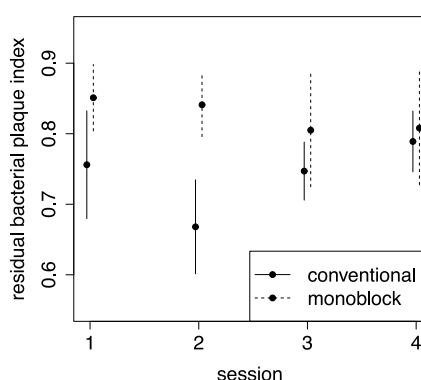
We did not identify evidence against the hypothesis that  $\gamma_{jd} = 1$ ,  $j = 0, 1$ ,  $d = 1, 2, 3, 4$ , under model (5) and (6) ( $p = 0.310$ ). The estimates of the parameters for this model are displayed in Table 3. Note that Nobre and Singer (2007) did not verify whether  $\gamma_{jd} = 1$ .

The conventional toothbrush seems more efficient than the monoblock toothbrush as suggested in Fig. 6.  $p$  values for interaction and main effects tests are presented in Table 4. The difference in



**Table 3** Estimates (Est.) and standard errors (SE) of parameters of LNLMM, GLMM, and GEE for the complete and reduced data sets assuming  $\gamma_{jd} = 1$ .

Parameter	LNLMM		GLMM		GEE	
	Complete data		Complete data		Complete data	
	Est.	SE	Est.	SE	Est.	SE
$\beta_{01}^*$	0.756	0.039	0.749	0.025	0.749	0.025
$\beta_{02}^*$	0.668	0.034	0.691	0.024	0.669	0.030
$\beta_{03}^*$	0.747	0.021	0.747	0.020	0.746	0.018
$\beta_{04}^*$	0.789	0.022	0.789	0.021	0.788	0.017
$\beta_{11}^*$	0.851	0.024	0.851	0.023	0.851	0.024
$\beta_{12}^*$	0.841	0.023	0.841	0.023	0.840	0.021
$\beta_{13}^*$	0.805	0.041	0.818	0.028	0.801	0.033
$\beta_{14}^*$	0.808	0.041	0.843	0.029	0.814	0.040
$\sigma_b^2$	0.005	0.002	0.004	0.002	0.003	0.002
$\tau_1^2$	0.033	0.007	0.014	0.003	0.013	0.003
$\tau_2^2$	0.007	0.002	0.008	0.002	0.008	0.002
$\alpha$					0.331	0.139
$\phi$					0.009	0.002
$\phi_1$					0.018	0.009
					0.247	0.087
					0.009	0.002
					0.006	0.003



**Figure 6** Estimated relative residual bacterial plaque index and corresponding 95% confidence intervals based on the LNLMM (complete data set).

**Table 4**  $p$  values for interaction and main effects tests for LNLMM, GLMM, and GEE.

Model	Interaction effect	Session effect		Toothbrush effect			
		Conventional	Monoblock	Session 1	Session 2	Session 3	Session 4
LNLMM	0.010	0.006	0.387	0.021	<0.001	0.266	0.840
GLMM	0.011	0.004	0.432	0.013	<0.001	0.173	0.597
GEE	0.009	0.004	0.544	0.004	<0.001	0.127	0.545

the expected efficiencies of the two types of toothbrush is not the same for all sessions ( $p = 0.010$ ): for the monoblock toothbrush it is constant along the sessions ( $p = 0.387$ ) but this is not so for the conventional toothbrush ( $p = 0.006$ ). The expected efficiencies are different for the monoblock and conventional toothbrushes in sessions 1 and 2 ( $p = 0.021$  and  $p < 0.001$ , respectively) and are similar in sessions 3 and 4 ( $p = 0.266$  and  $p = 0.840$ ). This suggests that the monoblock toothbrush is less efficient only in sessions 1 and 2 and is as efficient as the conventional toothbrush in sessions 3 and 4. In Nobre and Singer (2007), under the constant coefficient of variation assumption, the interaction effects were not significant.

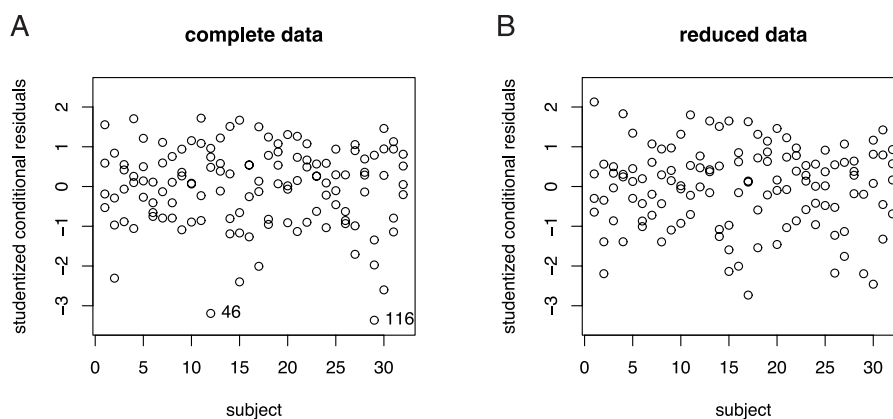
#### 4 Analysis via GLMM

In this section, we consider a GLMM based on a gamma distribution. As the LNLMM, this model is appropriate for a non-negative variable like the posttest bacterial plaque index and meets all the requirements set forth in Section 2. A positive correlation among the repeated measures is induced by the inclusion of random effects.

The typical gamma GLMM with a logarithmic link function for the posttest bacterial plaque indices,  $\mathbf{y}_{ij} = (y_{ij1}, \dots, y_{ij4})^\top$ , may be specified as

$$y_{ijd}|b_{ij} \sim \text{Gamma}(\mu_{ijd}, \phi), \quad (7)$$

$$\mu_{ij} = E(\mathbf{y}_{ij}|b_{ij}), \quad \log(\mu_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{1}_4 b_{ij}, \quad b_{ij} \sim N(0, \sigma_b^2), \quad (8)$$



**Figure 7** Studentized conditional residuals for the GLMM model versus subject indices. (A) Complete data. (B) Reduced data.

with  $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \dots, \mu_{ij4})^\top$  and  $\text{Var}(\mathbf{y}_{ij}) = \phi \mathbf{A}_{\boldsymbol{\mu}_{ij}}$ , where  $\mathbf{A}_{\boldsymbol{\mu}_{ij}} = \text{diag}(\mu_{ij1}^2, \mu_{ij2}^2, \mu_{ij3}^2, \mu_{ij4}^2)$ ,  $\phi$  is a dispersion parameter and  $\boldsymbol{\beta}$  and  $\mathbf{X}_{ij}$  are defined in (5). In this context, the coefficient of variation of the ratio of the posttest bacterial plaque index to the pretest bacterial plaque index is constant for all treatments. Since this disagrees with our previous conclusions, we consider an alternative GLMM, maintaining (8) and replacing (7) with

$$\text{Var}(\mathbf{y}_{ij}|b_{ij}) = \mathbf{A}_{\boldsymbol{\mu}_{ij}}^{1/2} \mathbf{W}_{ij} \mathbf{A}_{\boldsymbol{\mu}_{ij}}^{1/2}, \quad (9)$$

where, by taking  $\mathbf{W}_{ij} = \mathbf{R}_j$  as in (6), we include the heteroskedastic pattern identified in Fig. 3. If we let  $\mathbf{W}_{ij} = \phi \mathbf{I}$ , it follows that (9) is less restrictive than (7) since it does not require  $y_{ijd}|b_{ij}$  to follow a gamma distribution as pointed out by Jiang (2007).

We estimated the parameters in (8)–(9) using the restricted pseudo-likelihood method presented in Wolfinger and O’Connell (1993) and implemented in the PROC GLIMMIX of SAS (Littell et al., 2006). This method is based on a first-order Taylor approximation to  $\log(\mathbf{y}_{ij})$  around  $\boldsymbol{\mu}_{ij}$ . In order to maximize the restricted pseudo-likelihood function, a Newton-Raphson ridge optimization technique is considered.

Estimates of  $\beta_{jd}^* = E(y_{ijd}|x_{ijd} = 1) = E[E(y_{ijd}|b_i, x_{ijd} = 1)]$  as well as of the other parameters in the GLMM are presented in Table 2. Also, there is no evidence against the hypothesis  $\gamma_{jd} = 1$  ( $p = 0.441$ ) and estimates of the parameters of model (8) and (9) with  $\gamma_{jd} = 1$  are presented in Table 3.

The studentized conditional residuals in Fig. 7 present the same pattern as the corresponding residuals obtained for the LNLMM displayed in Fig. 4.

## 5 Analysis via GEE-based models

Here we analyze the pretest–posttest data via models where the marginal distribution is not completely specified; such models depend only on a dispersion parameter and on a vector of parameters ( $\boldsymbol{\theta}$ ) that indexes the mean  $\boldsymbol{\mu}_i(\boldsymbol{\theta})$  and the variance  $\mathbf{V}_i(\boldsymbol{\theta})$ . Estimates of the parameters are obtained as solutions to the GEE

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^K \frac{\partial \boldsymbol{\mu}_i^\top}{\partial \boldsymbol{\beta}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}, \quad (10)$$

where  $\Sigma_i = \phi \text{diag}(\mathbf{V}_i)^{1/2} \mathbf{W}(\alpha) \text{diag}(\mathbf{V}_i)^{1/2}$  and  $\mathbf{W}$  is a working covariance matrix that incorporates the within subject correlation. If we set  $V_i(\mu_i) = \mu_i^2$  and  $\mathbf{W} = \mathbf{I}$ , the solution is equivalent to a GLM with a gamma distribution and a log-link function.

For the sake of comparison, we consider a GEE-based model for which

$$\boldsymbol{\mu}_{ij} = E(\mathbf{y}_{ij}), \quad \log(\boldsymbol{\mu}_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}, \quad \text{Corr}(y_{ijd}, y_{ijd'}) = \alpha, d \neq d', \quad (11)$$

where  $\mathbf{X}$  and  $\boldsymbol{\beta}$  are defined in (5) and  $\phi_{jd}$  are dispersion parameters. To capture the behavior of the coefficients of variation identified in Fig. 3, the dispersion parameter may depend on the type of toothbrush ( $j$ ) and session ( $d$ ), that is,

$$\phi_{jd} = \begin{cases} \phi, & \text{if } (j = 0 \text{ and } d = 3 \text{ or } 4) \text{ or } (j = 1 \text{ and } d = 1 \text{ or } 2) \\ \phi + \phi_1, & \text{otherwise.} \end{cases} \quad (12)$$

This model was fitted via the R package `geepack` (Halekoh et al., 2006) using GEE that allows for different covariates and even different link functions and estimating equations for the mean, scale, and correlations as considered in Yan and Fine (2004). The estimates of  $\beta_{jd}^*$  and of the other parameters of model (11) and (12) are presented in Table 2. Again, no evidence against the hypothesis  $\gamma_{jd} = 1$  ( $p = 0.481$ ) was detected and the estimates under this restriction are presented in Table 3.

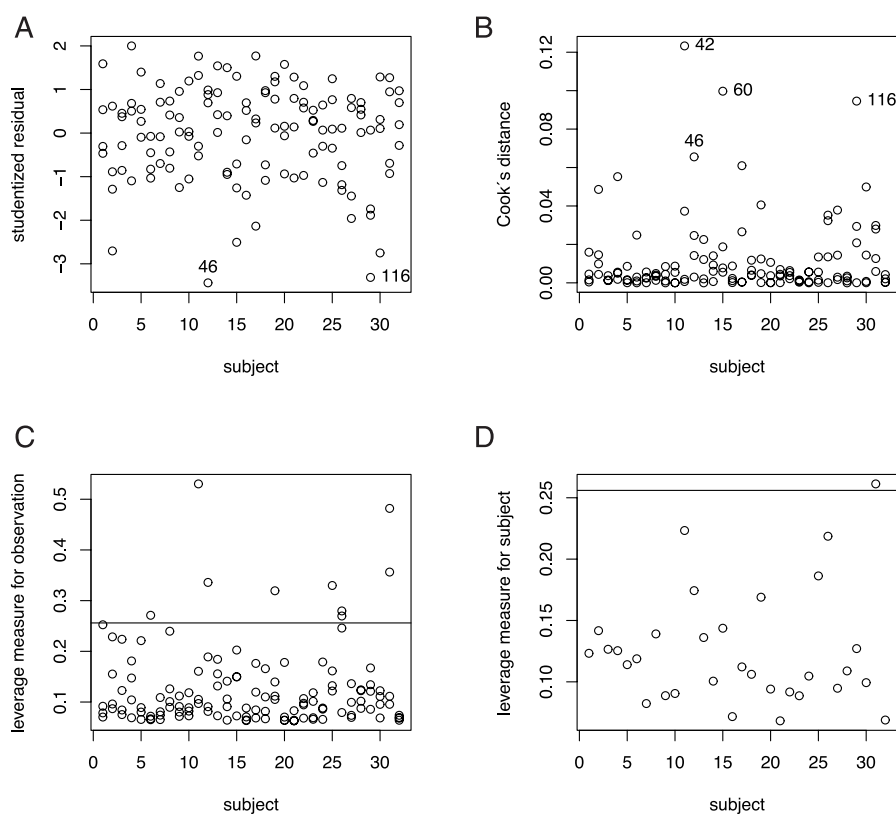
Plots of the studentized residuals are discussed in Venezuela et al. (2007, 2011) as well as Vens and Ziegler (2011). These residuals are presented in Fig. 8 and their patterns are similar to those in Fig. 4. These authors also show how to compute Cook's distance and a leverage measure in this setting. These quantities are also presented in Fig. 8. The outlying observations are the same identified previously. Although, these observations do not present high leverage (Fig. 8B), the corresponding Cook's distances are large for these outliers. Moreover, excluding them, no other observation is identified as an outlier.

## 6 Discussion

We considered three families of linear models that are sufficiently flexible to accommodate different characteristics of longitudinal data. In particular, they may account for nonconstant coefficients of variation, a feature that may justify differences generated by underlying log-normal and gamma distributions, as discussed in Wiens (1999).

Differences between estimates obtained under marginal and conditional models may occur due to noncomparable parameters, as indicated by Lee and Nelder (2004) and Fitzmaurice et al. (2008). In our example, the main goal was to evaluate the efficiency of two types of toothbrushes along 45 days. This may not be carried out by a direct comparison of the estimates of  $\beta_{jd}$  in models (5) and (6), (8) and (9), and (11) and (12), because such parameters have different interpretations. Assuming the correct specification of all models, the estimates of the original parameters  $\beta_{jd}$  were smaller for LNLMM than for GLMM- or GEE-based models. To bypass this problem, we compared the population averaged effects that correspond to the relative residual bacterial plaque indices,  $\beta_{jd}^* = E(y_{jd} | x_{jd} = 1)$ . For both LNLMM and GLMM, computation of this marginal expectation requires integration of the random effects in  $E[\mathbf{Y}_i | \mathbf{b}_i]$ , while GEE-based models allow a direct estimation, an advantage reported by Serroyen et al. (2009), especially for nonlinear mixed models.

In general, the estimates of the fixed parameters and their corresponding standard errors obtained under LNLMM-, GLMM-, or GEE-based models are very close to each other, both with or without the exclusion of the outliers. The estimates of the variance components are also similar for LNLMM and GLMM. The estimate of  $\tau_1^2$  is the most affected by the outliers because it is related to the extra-variance for the treatments where the outliers occurred. For the marginal model (GEE-based),



**Figure 8** Residual analysis for the GEE-based model with complete data. (A) Studentized conditional residuals. (B) Cook's distance. (C) Leverage measure for each observation. (D) Leverage measure for each subject.

the estimated correlation of repeated measures is smaller for the data without outliers and so is its standard error. The conclusions of all tests are the same for the three approaches, although the  $p$  values are in general smaller for GEE-based models (see Table 4).

For random-effects models, like LNLMM and GLMM, another important issue is the robustness of estimation of fixed effects with respect to misspecification of the distribution of the random effects as discussed in Fitzmaurice et al. (2008). Verbeke and Lesaffre (1997) showed that ML estimators of fixed effects obtained for a linear mixed model under Gaussian distributed random effects are consistent even when the random effects are non-Gaussian. Also, restricted ML estimators of the variance components are consistent; this is important because they are used to obtain estimates of both  $\beta_{jd}^*$  and of its variance. Moreover, Verbeke and Lesaffre (1996) also comment that it is not possible to assess the distribution of the random effects by analyzing the predicted  $\hat{b}_i$ , which makes estimation robustness even more important. For GLMM, Litière et al. (2008) conclude that misspecification of the distribution of random effects may produce biased ML estimators of fixed effects but that this bias is small when the variability of random effects is small; however, as the estimators of this variability are severely biased, it is not possible to evaluate whether the biases of fixed effect estimators are negligible or not. An additional advantage of GEE-based models is that GEE provide consistent estimators of fixed effects if the corresponding model is correctly specified regardless of the correlation and scale structures (Liang and Zeger, 1986). Also, when the scale parameter depends on covariates, the corresponding estimators are consistent if the fixed effects and scale structures are correctly specified

**Table 5** Diagnostic tools and references.

Model	Diagnostic tools	Reference
LNLMM	Conditional and marginal residuals and empirical best linear unbiased estimator (EBLUP)	Nobre and Singer (2007)
	Generalized leverage matrix	Nobre and Singer (2011)
	Cook's distance	Tan et al. (2001)
GLMM	Goodness-of-fit tests	Vonesh et al. (1996)
	Cook's distance	Xiang et al. (2002)
	Influence	Zhu and Lee (2003)
	Goodness-of-fit for random effect	Waagepetersen (2006)
GEE	Standardized residuals, projection matrices, Cook's distance	Venezuela et al. (2007)
	Local influence	Venezuela et al. (2011)
	Standardized residuals, projection matrices, Cook's distance	Vens and Ziegler (2011)

(Yan and Fine, 2004). Therefore, estimates of the fixed effects are more robust under GEE-based models and LNLMM than under GLMM if the covariance structure is misspecified.

Diagnostic tools are well developed for LNMM (see Nobre and Singer [2007], e.g.). For GLMM, Vonesh et al. (1996) propose a goodness-of-fit statistic to evaluate the adequacy of an assumed mean and covariance structure and an approximate pseudo-likelihood test for the adequacy of the covariance structure; Xiang et al. (2002) use Cook's distance for clustered data to identify influential clusters and Zhu and Lee (2003) propose a different method to measure influence for GLMM. Also, Waagepetersen (2006) assesses the goodness-of-fit of the random effects distribution using simulation. Nevertheless, more research is needed for residual analysis for GLMM as already indicated in Dean and Nielsen (2007). Diagnostic tools for the GEE are presented, for example, in Venezuela et al. (2007), Venezuela et al. (2011), and Vens and Ziegler (2011) to detect influential and outlying observations using the projection (hat) matrix, Cook's distance, standardized residuals, and half-normal plots of absolute residuals with simulated envelopes, the latter only for constant scale parameter as proposed in Park and Shin (1998). Some diagnostic tools and references are presented in Table 5.

Other estimation issues involving the GEE method are related to missing data problems and the estimation of the correlation matrix. For example, Lu et al. (2009) show that fixed-effect estimators of LMM are biased under departures from normality in the presence of missing data. Also, based on simulation, they conclude that other (weighted and augmented weighted) robust estimators based on GEE provide valid inference for skewed non-Gaussian data when missing data follows a missing at random pattern. Sun et al. (2009) compare estimation methods for the correlation in the GEE framework and conclude that the degree of imbalance and variability in the temporal spacing of measurements, the value of the correlation, and the type of outcome affect the choice of the best method.

A practical issue to be addressed is the availability of computational software to fit the families of models under consideration. Although we used R only to fit GEE-based models, there are libraries to fit GLMM (package MASS with the command `glmPQL` using penalized quasi-likelihood) and LNMM (`lme4` and `nlme`). Table 6 presents procedures of SAS and R packages to estimate the studied models.

If on the one hand estimates obtained under LNMM are robust to misspecification of random effects distributions, then on the other hand GEE-based models may be preferable for computational reasons, when the main goal is only to estimate and compare marginal expected responses. LNMM may be preferable to estimate subject-specific expected responses since a variety of diagnostic analyses are available for this family. To help users with respect to the choice of appropriate families of models, estimation procedures, software, and diagnostic tools, we summarize different characteristics for the three approaches in Table 5. Other estimation methods are available, for example, in Fitzmaurice et al. (2008).

**Table 6** Available statistical packages using SAS and R and estimation methods used in this study.

Model	SAS PROC	R packages	Estimation method
LNLM	MIXED	lme4 and nlme	REML (Littell et al., 2006)
GLMM	GLIMMIX	MASS (glmmPQL)	Pseudo likelihood (Wolfinger and O'Connell, 1993)
GEE	GENMOD	geepack	GEE (Halekoh et al., 2006)

Finally, we mention that because of the relatively small (but not uncommon in studies of this type) sample size (32 children with four observations/child), the conclusions based on asymptotic results should be viewed with caution. A good practice in such cases is to fit different models to the data, use different diagnostic tools to detect possible outliers, and verify whether the results are coherent. In our case, the three models generated similar results even when the possible outliers were included, suggesting that the conclusion seems pertinent.

**Acknowledgments** This research was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil. The authors also acknowledge the constructive comments and suggestions of the anonymous referees and the associate editor.

#### Conflict of interest

*The authors have declared no conflict of interest.*

#### References

- Aoki, R., Achcar, J. A., Bolfarine, H. and Singer, J. M. (2003). Bayesian analysis of null intercept errors-in-variables regression for pretest/posttest data. *Journal of Applied Statistics* **30**, 3–12.
- Dean, C. B. and Nielsen, J. D. (2007). Generalized linear mixed models: a review and some extensions. *Lifetime Data Analysis* **13**, 497–512.
- Demidenko, E. (2004). *Mixed Models Theory and Applications*. Wiley, New York.
- Diggle, P., Liang, K. and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.
- Fitzmaurice, G. M., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*. Chapman & Hall/CRC, Boca Raton, FL.
- Fitzmaurice, G. M., Laird, N. and Ware, J. (2004). *Applied Longitudinal Analysis*. Wiley, New York.
- Guerin, L. and Stroup, W. W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Proceedings of the 12th Annual Conference on Applied Statistics in Agriculture*, 170–203. Manhattan: Kansas State University.
- Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician* **60**, 19–26.
- Halekoh, U., Hojsgaard, S. and Yan, J. (2006). The R Package geepack for generalized estimating equations. *Journal of Statistical Software* **15**, 1–11.
- Hardin, J. W. and Hilbe, J. M. (2002). *Generalized Estimating Equations*. Chapman and Hall/CRC, Boca Raton, FL.
- Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.
- Jiang, J. and Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhyā* **65**, 23–42.
- Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.
- Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science* **19**, 219–238.
- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.



- Litière, S., Alonso, A. and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine* **27**, 3125–3144.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Institute Inc, Cary, NC.
- Lu, N., Tang, W., He, H., Yu, Q., Crits-Christoph, P., Zhang, H. and Tu, X. (2009). On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. *Biometrical Journal* **51**, 627–643.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.
- Nobre, J. S. and Singer, J. M. (2007). Residual analysis for linear mixed models. *Biometrical Journal* **49**, 1–13.
- Nobre, J. S. and Singer, J. M. (2011). Leverage analysis for linear mixed models. *Journal of Applied Statistics* **38**, 1063–1072.
- Park, C. G. and Shin, D. S. (1998). An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation* **61**, 127–139.
- Serroyen, J., Molenberghs, G., Verbeke, G. and Davidian, M. (2009). Nonlinear models for longitudinal data. *The American Statistician* **63**, 378–388.
- Singer, J. M. and Andrade, D. F. (1997). Regression models for the analysis of pretest/posttest data. *Biometrics* **53**, 729–735.
- Singer, J. M., Nobre, J. S. and Sef, H. C. (2002). Regression models for pretest/posttest data in blocks. *Statistical Modelling* **4**, 1–15.
- Song, P. X. K. (2007). *Correlated Data Analysis—Modelling, Analytics and Applications*. Springer, New York.
- Sun, W., Shults, J. and Leonard, M. (2009). A Note on the use of unbiased estimating equations to estimate correlation in analysis of longitudinal trials. *Biometrical Journal* **51**, 5–18.
- Tan, F. E. S., Ouwens, M. J. N. and Berger, M. P. F. (2001). Detection of influential observations in longitudinal mixed effects regression models. *Journal of the Royal Statistical Society: Series D (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 generalized estimating equations. *Computational Statistics and Data Analysis* **55**, 1867–1883.
- Vens, M. and Ziegler, A. (2011). Generalized estimating equations and regression diagnostics for longitudinal controlled clinical trials: a case study. *Computational Statistics and Data Analysis*. Available at <http://www.sciencedirect.com/science/article/pii/S0167947311001393>.
- Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217–221.
- Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis* **53**, 541–556.
- Vonesh, E. F., Chinchilli, V. M. and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics* **52**, 572–587.
- Waagepetersen, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scandinavian Journal of Statistics* **33**, 721–731.
- Wiens, B. L. (1999). When log-normal and gamma models give different results: a case study. *The American Statistician* **53**, 89–93.
- Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computing and Simulation* **48**, 233–243.
- Xiang, L., Tse, S. K. and Lee, A. (2002). Influence diagnostics for generalized linear mixed models: applications to clustered data. *Computational Statistics and Data Analysis* **40**, 759–774.
- Yan, J. and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine* **23**, 859–874.
- Zhu, H. T. and Lee, S. Y. (2003). Local influence for generalized linear mixed models. *The Canadian Journal of Statistics* **31**, 293–309.