Computational Statistics and Data Analysis 56 (2012) 1405–1416

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Residual analysis of linear mixed models using a simulation approach

André Schützenmeister, Hans-Peter Piepho *

Bioinformatics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 11 November 2010
Received in revised form 29 September 2011
Accepted 3 November 2011
Available online 17 November 2011

Keywords:
Studentized residual
Simultaneous tolerance band
Simultaneous tolerance interval
Diagnostic plot
Conditional residual
Empirical size

ABSTRACT

In the framework of the general linear model, residuals are routinely used to check model assumptions, such as homoscedasticity, normality, and linearity of effects. Residuals can also be employed to detect possible outliers. Various types of residuals may be defined for linear mixed models. It is shown how residual plots can be used to check model assumptions by comparing empirical residual distributions with appropriate null distributions based on a parametric bootstrap approach. This allows constructing simultaneous tolerance bounds, which helps in assessing the normality and homoscedasticity of residuals of linear mixed models, identifying possible outliers and interpreting residual plots. The usefulness of this method is demonstrated by applying it to several previously published datasets.

1. Introduction

Linear mixed models (LMM) provide a flexible framework for the analysis of various types of data. They are a natural extension of linear models (LM), where only a single random term is included, the residual error. LMMs allow specification of more than one random term. This is a useful feature, since it is often more convenient to think of an effect as coming from a specific normal distribution rather than having a fixed value. Many experimental designs require fitting more than one random error term (Cochran and Cox, 1957). Robinson (1991) gives an excellent introduction to the estimation of random effects and discusses its various fields of application. LMMs originated from estimation of genetic merits in animal breeding, where they were first introduced by Henderson (1950). The great flexibility of random effects estimation, commonly referred to as prediction, was soon exploited in various application domains including, for example, estimating ore deposits using a procedure known as kriging, insurance credibility theory, and digital image processing (Robinson, 1991). With the availability of molecular marker data, LMMs are now widely used for genomic selection in animal and plant breeding (Meuwissen et al., 2001; Piepho, 2009). Another important area of application is in the analysis of repeated measures and longitudinal data (Verbeke and Molenberghs, 2000).

A key assumption when making inferences using LMMs is that the residual errors and random effects are normally distributed. Lange and Ryan (1989) investigated the normality of random effects in LMMs for repeated measures data (longitudinal data). They proposed generalized weighted normal plots, with the weights chosen to reflect the differing sampling variances of the estimated random effects. Verbeke and Molenberghs (2000, p. 89), pointed out the inadequacy of these plots as they cannot differentiate between a wrong choice of covariates and wrong distributional assumptions on the error terms or the random effects.

Recently, Gumedze et al. (2010) extended a variance shift outlier model (VSOM; (Thompson, 1985)) to LMMs. The rationale of a VSOM is to add an extra random effect to the model, which accounts for extra variability introduced by a specific

* Corresponding author. Tel.: +49 711 459 22386; fax: +49 711 459 22297.
E-mail address: piepho@uni-hohenheim.de (H.-P. Piepho).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.11.006

observation. Thus, numerical values are assigned to observations to quantify the amount of inflated variance attributable to each observation. These can then be used to classify each observation as an outlier or not. Inflated variance estimates may be employed to assign weights to observations when fitting the LMM. The VSOM approach can also be applied to groups of observations, which allows an assessment of single random effects. Gumedze et al. stress that a VSOM has many important benefits compared to case-deletion approaches (Cook and Weisberg, 1982), which are also available in LMMs (Christensen et al., 1992; Haslett and Dillane, 2004), especially when groups of outliers are being considered. Shi and Chen (2012) show that the equivalence of three common approaches to identifying influential observations and outliers in a large class of LMMs (mean-shift, deletion, replacement) depends on the use of either Maximum Likelihood (ML) or restricted ML (REML) employed to estimate variance components. As well, Shi and Huang (2011) present a step-wise local influence analysis method for detecting influential observations that is applicable to LMMs and able to reveal the presence of masking effects, a common problem in identifying influential observations. Longford (2001) provides an excellent discussion about outliers and proposes simulation-based diagnostics for random coefficient models. He proposes the parametric bootstrap (Efron and Tibshirani, 1993) to dispense with asymptotic theory and base inference on an approximative null distribution of an appropriate diagnostic feature, which can be a statistic or a graphical feature. Nobre and Singer (2007) explore the residual analysis of LMMs for repeated measures data. They also review different types of residuals, which arise in the analysis of LMMs and present some theory. Nobre and Singer also summarize areas of application for each type of residual defined for LMMs, e.g. checking linearity of effects, assessing the covariance structure for individual subjects, checking for outliers, and assessing normality and homoscedasticity of residuals. Recently, Huang (2011) extended a diagnosing-method for the misspecification of random-effects in generalized LMMs (GLMM) suitable for binary response and based on data-coarsening to a much wider class of GLMMs with non-binary response.

The assumed normality of residual errors and normality of random effects may be assessed with quantile–quantile (QQ) plots (Pinheiro and Bates, 2000). QQ-plots are very helpful, but their use always involves some unavoidable subjectivity, since it is not generally obvious whether the observed pattern is acceptable or not. Therefore, it is desirable to add tolerance bounds to such plots, to reflect the null distribution for a specific diagnostic feature, and enhance objectivity in interpreting these plots.

In this paper, we are mainly concerned with assessing normality and homoscedasticity of various types of residuals in an LMM. Both applications provide a means to identify possibly outlying observations. Our approach is based on the parametric bootstrap, which allows generation of approximative null distributions of graphical features. The manuscript is organized as follows. We start with an example to demonstrate the general scope of our method. In Sections 3 and 4, we present the underlying theory, and proceed in Section 5 by exemplifying our method using two published datasets. In Section 6, we present a small simulation study to infer whether this approach maintains the expected error rate.

2. Motivating example

Nobre and Singer (2007) illustrated residual analysis for LMMs using data to compare two types of toothbrushes, a low-cost mono-block toothbrush and a conventional toothbrush. The main interest lay in the maintenance of the capacity to remove bacterial plaque under daily use. This dataset consisted of 32 children, aged 6–8, of which one half used the conventional toothbrush, while the other half used the low-cost toothbrush. In four sessions, bacterial plaque indices were evaluated before and after using the respective toothbrushes. Obviously, the data comprise repeated measures on the same experimental unit/subject (child). Nobre and Singer (2007) used the LMM

$$\log(y_{ijd}) = \alpha_j + \beta \log(x_{ijd}) + b_i + e_{ijd}, \tag{1}$$

where y_{ijd} is the post-, x_{ijd} is the pre-treatment bacterial plaque-index of the i-th subject, in session d, using the j-th type of toothbrush, α_j is the fixed effect of the j-th type of toothbrush, β is a fixed regression coefficient, and $b_i \sim N(0; \sigma_s^2)$ and $e_{ijd} \sim N(0; \sigma_e^2)$ are independent random variables, where the former corresponds to the random subject effect and the latter to random measurement error.

Fig. 1a depicts the QQ-plot of studentized conditional residuals (CR, see Section 3), i.e. the studentized estimates of the residual errors ($\hat{e}^*_{ijd}$), well known from residual analysis of LMs. The problem for this type of plot is the difficulty of assessing whether the plot is indicative of a departure from normality and/or whether there are possible outliers. These problems are even more evident when observation 2 of subject 12 (12.2) and observation 4 of subject 29 (29.4) are removed, which results in a QQ-plot that for some observers might not raise concerns about non-normality at the first glance, while others would still see some unacceptable curvature in the plotted residuals (Fig. 1c). The two offending observations (12.2, 29.4) were identified and classified as outlying observations by Nobre and Singer (2007).

The interpretation of QQ-plots and residual plots, as shown in Fig. 1, greatly benefits from sketching a tolerance area/interval, which represents the expectation under a specific assumption, e.g. normality (Fig. 1a, c) or homoscedasticity (Fig. 1b, d). The corresponding bounds define a region, where a pre-defined proportion of the whole population of a diagnostic feature falls within, e.g. studentized CRs (Section 3). Atkinson (1981, 1985) suggests computing envelopes in half-normal or QQ-plots in the context of linear regression, which are basically simulation-based point-wise tolerance intervals (TI) for each order statistic of the residual vector. These envelopes facilitate the interpretation of half-normal or QQ-plots significantly, in particular for less experienced users. Schützenmeister et al. (2012) suggested $100(1-\alpha)\%$ simultaneous

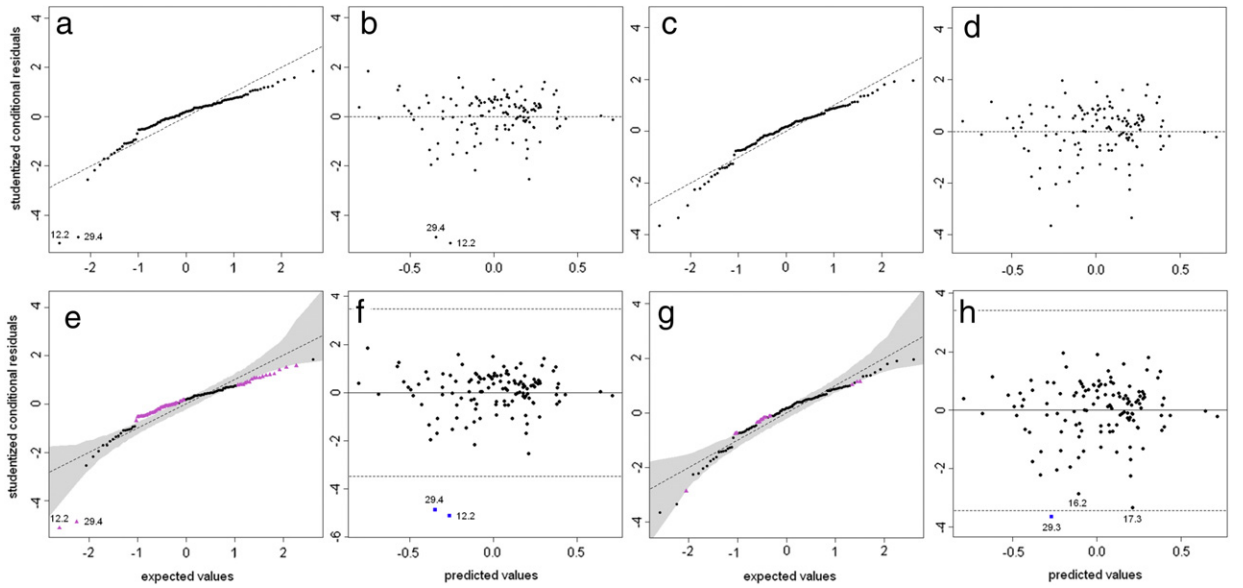

Fig. 1. (a): QQ-plot of studentized conditional residuals (CR) for the complete toothbrush data and model (1); (b): Residual plot for the complete toothbrush data for model (1); (c): QQ-plot of CRs for model (1), where observations 12.2 and 29.4 were removed; (d): Residual plot of the toothbrush data for model (1), where observation 12.2 and 29.2 were removed; (e–h): Plots correspond to the plots above, where simultaneous tolerance bands (STB) were added to the QQ-plots (e, g), and simultaneous tolerance intervals (STI) were added to the plots of studentized conditional residuals vs. predicted values (f, h).

tolerance bounds for residuals in LMs, which take the multiplicity problem into account. Due to its conceptual simplicity, this approach can readily be extended to the various types of residuals defined for LMMs, as will be illustrated in this paper.

Fig. 1b and d depict residual plots useful for identifying possible dependence of the residual variance on predicted values and can be used to identify possible outliers. As for the QQ-plots, it may be hard to decide conclusively whether the assumptions of normality and/or homoscedasticity are met. In the lower row of Fig. 1 (e–h), the same plots are depicted as in the upper row (a–d), but with simultaneous tolerance bands (STB) and intervals (STI) added. There is some indication that the assumptions in question are still violated even after removal of the two outliers identified by Nobre and Singer (2007). We also analyzed the data without transforming x and y. This removed the curvature problem, which caused mid-range CRs to fall outside the $100(1-\alpha)\%$ STB (results not shown). It is noteworthy, however, that Singer and Andrade (1997) and Singer et al. (2004) identified subject-matter related characteristics of the data that favor a logarithmic transformation, so the need to transform these data may be somewhat controversial.

3. Mixed models and residuals

The general specification of LMMs in standard matrix notation can be written as

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

Here, $\mathbf{y}$ is an $(n \times 1)$ vector of response variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects linked to each observation via the $(n \times p)$ design/model matrix $\mathbf{X}$, where $p = \text{rank}(\mathbf{X})$, $\mathbf{Z}$ is the $(n \times q)$ design/model matrix linking the q random effects in $\mathbf{b}$ to each observation, and $\mathbf{b}$ and $\mathbf{e}$ are independent random variates, which are normally distributed with

$$E\begin{bmatrix}\mathbf{b}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \quad \text{and} \quad \text{Var}\begin{bmatrix}\mathbf{b}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}. \tag{3}$$

Matrices $\mathbf{G}$ and $\mathbf{R}$ are covariance structures of random effects $\mathbf{b}$ and residual errors $\mathbf{e}$, respectively, which form, upon incorporating matrix $\mathbf{Z}$, the variance of $\mathbf{y}$, $\text{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$, which has expectation $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. A key assumption in the analysis of LMMs is that variances and covariances obey the structure shown in (3) and that $\mathbf{V}$ is positive-definite.

Nobre and Singer (2007) reviewed three types of residuals for LMMs useful for three types of residual analysis, namely marginal residuals $\hat{\boldsymbol{\varepsilon}}$, $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Z}\hat{\mathbf{b}} + \hat{\mathbf{e}}$, conditional residuals $\hat{\mathbf{e}}$ (CR), $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}$, and best linear unbiased predictions (BLUPs) $\mathbf{Z}\hat{\mathbf{b}}$. Marginal residuals are useful for checking for linearity of effects and the covariance structure of $\mathbf{V}$, whereas CRs can be used to check for outlying observations, homoscedasticity and normality of residual errors. $\mathbf{Z}\hat{\mathbf{b}}$ can be employed to check for outlying subjects, the random effects covariance structure $\mathbf{G}$, and the normality of random effects $\mathbf{b}$ (Nobre and Singer, 2007). If a model has more than one random effect besides the residual error, it is often useful to check the random effects separately (see e.g. Pinheiro and Bates (2000, p. 189)). The advantage of using $\mathbf{Z}\hat{\mathbf{b}}$ with multiple random effects (besides the residual error) is that the matrix $\mathbf{Z}$ suitably links several random effects, which may add up

to yield unusually large deviates (as absolute values) for single observations (rows in $\mathbf{Z}$). In contrast, the random effects estimates/predictions themselves can be completely inconspicuous if checked separately.

Nobre and Singer (2007) note that CRs as well as $\mathbf{Z}\hat{\mathbf{b}}$ are not pure, which means, according to Hilden-Minton (1995), that they are not independent of other types of errors. CRs are confounded with the vector of random effects $\mathbf{b}$, and $\mathbf{Z}\hat{\mathbf{b}}$ are confounded with $\mathbf{e}$ (Hilden-Minton, 1995; Nobre and Singer, 2007). Both, Hilden-Minton and Nobre and Singer propose a linear transformation to obtain $(n - r)$ least-confounded (Nobre and Singer, 2007) or unconfounded residuals (Hilden-Minton, 1995) corresponding to errors $\mathbf{e}$, where $r = \text{rank}[\mathbf{X}|\mathbf{Z}]$. This approach is similar to orthogonal residuals for LMs (reviewed in (Cook and Weisberg, 1982; Seber, 1977)).

There are two distinct problems with any linear transformation applied to residuals. First, the direct interpretation of a residual point in e.g. a QQ-plot gets blurred (Cook and Weisberg, 1982), since linearly transformed residuals do not correspond to individual observations any more. Second, each type of residual vector represents estimates of the unobservable, underlying true errors, as a linear combination. We think that any additional linear transformation of residuals may amplify the *supernormality* effect, which occurs when a set of estimates looks more normal than the underlying effects actually are (Atkinson, 1985). For example, Verbeke and Lesaffre (1996) showed that BLUPs can look normal even in cases, where the underlying distribution is non-normal.

Hilden-Minton (1995) points out that the space of least-/unconfounded residuals is identical to the residual space of the fixed effects analysis of the original LMM (the space orthogonal to $\mathbf{R}^{-1/2}[\mathbf{X}|\mathbf{Z}]$), where fixed effects analysis means taking random effects of the LMM as fixed. Moreover, Hilden-Minton points out that a fixed effects analysis can be used to diagnose specific parts of an LMM. This allows computation of the associated simultaneous tolerance bounds by simulation to any desired accuracy, since studentization of the estimated residuals in the fixed effects analysis makes them a pivotal quantity (Cox and Hinkley, 1974; Dufour et al., 1998; Schützenmeister et al., 2012). The coverage of these bounds becomes exact for $N \to \infty$, where N is the number of simulations. Unfortunately, the pivotal property does not carry over to studentized residuals of LMMs because of the confounding which takes place. We propose to tackle the problem of confounding by using a fixed effects analysis of the LMM, which then becomes an ordinary LM. Inspecting the studentized residuals of the LM is not influenced by any confounding of the random terms (random effects and errors). This allows an assessment of the normality of conditional errors, retains the connection to individual observations, and is not suspected of introducing further *supernormality*. We will exemplify this in the following sections.

The variance of the estimated CRs is $\text{Var}(\hat{\mathbf{e}}) = \mathbf{P} = \mathbf{RQR}$, where $\mathbf{Q} = \mathbf{V}^{-1}(\mathbf{I} - \mathbf{H})$, $\mathbf{H} = \mathbf{XT}$ (hat matrix), and $\mathbf{T} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. In practice, parameters in $\mathbf{R}$ and $\mathbf{G}$ need to be replaced by their estimates in order to estimate matrix $\mathbf{Q}$. Note, that matrix $\mathbf{T}$ can directly be used to obtain the generalized least squares estimate of the fixed effect parameter vector $\boldsymbol{\beta}$ (BLUE) as $\hat{\boldsymbol{\beta}} = \mathbf{Ty}$, thus $\hat{\mathbf{y}} = \mathbf{Hy}$. We will apply studentization for CRs. The k-th CR can be studentized as

$$\hat{e}_k^* = \frac{\hat{e}_k}{\sqrt{\hat{p}_{kk}}}, \tag{4}$$

where $\hat{p}_{kk}$ is an estimate of p_{kk}, the k-th diagonal element of matrix $\mathbf{P}$. The diagonal elements of $\mathbf{P}$ are functions of the joint leverage of fixed effects and random effects, thus constituting a generalization of studentized residuals (Nobre and Singer, 2007). A studentized version of the l-th estimated random effect $\hat{b}_l$ can be computed as

$$\hat{b}_l^* = \frac{\hat{b}_l}{\sqrt{\hat{o}_{ll}}}, \tag{5}$$

where $\hat{o}_{ll}$ is an estimate of o_{ll}, the l-th diagonal element of matrix $O = \text{Var}(\hat{\mathbf{b}}) = \mathbf{GZ}^T\mathbf{QZG}$ (Laird and Ware, 1982; Searle et al., 1992).

4. The simulation approach

The basic idea of our approach is to generate appropriate null distributions of residual plots (graphical features) by Monte Carlo simulation and to compare observed residuals and random effects to these. Specifically, we simulate N datasets under the null hypothesis of normality of conditional errors $\mathbf{e}$ and random effects $\mathbf{b}$, with covariance matrices $\mathbf{R}$ and $\mathbf{G}$, refit the specified model, and extract the different types of residuals.

Each vector of simulated data $\mathbf{y}_{sim}$, comprising n observations, is constructed as

$$\mathbf{y}_{sim} = \mathbf{Zb}_{sim} + \mathbf{e}_{sim} = [\mathbf{Z}|\mathbf{I}_n] \begin{bmatrix} \mathbf{b}_{sim} \\ \mathbf{e}_{sim} \end{bmatrix}, \tag{6}$$

where $\mathbf{b}_{sim}$ and $\mathbf{e}_{sim}$ have to be simulated, using the estimates of variance components, and $\mathbf{I}_n$ is an identity matrix of size n. An alternative and sometimes more convenient way to obtain $\mathbf{y}_{sim}$ is to use a Cholesky decomposition of $\mathbf{V}$, $\mathbf{V} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. A simulated vector $\mathbf{y}_{sim}$ with covariance $\mathbf{V}$ can be computed as

$$\mathbf{y}_{sim} = \boldsymbol{\Gamma}\mathbf{z}, \tag{7}$$

where $\mathbf{z}$ is a vector of independent standard normal deviates of size n. In fact, this is a classical parametric bootstrap approach (Efron and Tibshirani, 1993), which is obvious from (6). One does not have to include the fixed effects part of the model in the simulation, which becomes evident by looking at the construction of the random part of the model, the marginal residuals:

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\mathbf{T}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{X}\mathbf{T})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}) \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}) \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1})(\mathbf{Z}\mathbf{b} + \mathbf{e}).\end{aligned}$$

The last equality follows from $\mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ and can be simplified to $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})(\mathbf{Z}\mathbf{b} + \mathbf{e})$. Addition of $\mathbf{X}\hat{\boldsymbol{\beta}}$ to the simulated random part does not change the results, when the original LMM is refitted to simulated data. Invariance with respect to the value of $\boldsymbol{\beta}$ still holds when $\mathbf{V}$ is replaced by an estimator (Kackar and Harville, 1984). In case of $\mathbf{b}_{sim}$, there might be non-zero covariances involved, which have to be accounted for in the simulation process. For example, for longitudinal data, where repeated measures were taken over time on the same subject, and a random regression model is fitted, the random intercept and random slope must be allowed to have non-zero covariance in order to maintain invariance to scale-shift transformations of the covariate (see Section 5.2).

Note that simulation of null distributions regarding graphical features employs estimates of covariance matrices $\mathbf{R}$ and $\mathbf{G}$, which themselves can be influenced by outlying observations causing a masking effect. Our proposed method assumes that there are only few outliers, which should stand out from the remaining residual points. These are hopefully identified by employing simultaneous tolerance bounds. The amount by which such outlying observations influence estimates of $\mathbf{R}$ and $\mathbf{G}$ could be quantified by applying a VSOM (Gumedze et al., 2010). If there are observations which severely influence estimation of the residual variance, and hence of $\mathbf{R}$, they should be carefully examined.

We make use of two types of simultaneous tolerance bounds, when assessing normality and homoscedasticity of LMM residuals. Consider, without loss of generality, the j-th simulated vector $\tilde{\mathbf{e}}_j \;\; (j = 1, \ldots, N)$ of CRs, studentized according to (4), whose order statistics are denoted as $\tilde{e}_{(j,1)} \leq \tilde{e}_{(j,2)} \leq, \ldots, \leq \tilde{e}_{(j,n)}$ stored in the j-th row of a $(N \times n)$ matrix $\mathbf{S}$. A simultaneous tolerance band (STB) for the observed vector of studentized CRs $\tilde{\mathbf{e}}_{obs}$ can be computed from $\mathbf{S}$. It is formed by n point-wise tolerance intervals for each individual order statistic $\tilde{\mathbf{e}}_{(obs;i)}$, $(i = 1, \ldots, n)$. The (local) tolerance level γ used for each of the n intervals is chosen such that at least $100(1-\alpha)\%$ of all simulated residual vectors are simultaneously covered by the bounds of the local intervals, i.e. at most $100\alpha\%$ of these vectors violate these bounds. Thus, the STB has coverage greater than or equal to $100(1-\alpha)\%$, which corresponds to multiplicity corrected bounds of the n local intervals.

Schützenmeister et al. (2012) sketch a quantile-based algorithm for the computation of the local tolerance level γ. It can be improved in terms of computation time using a rank-based approach. The basic idea is to iterate over rows of the $(N \times n)$ matrix $\mathbf{S}$ and to remove those rows, which have at least one extreme studentized residual. This is done until αN rows have been removed, i.e. there are $(1-\alpha)N$ rows remaining. Then, the minimum and maximum values for each column (order statistic) define the local (point-wise) bounds of the $100(1-\alpha)\%$ STB.

This algorithm makes use of three $(N \times n)$ matrices. Matrix $\mathbf{S}$ is defined as shown above, while matrix $\mathbf{D}$ contains absolute, standardized values of matrix $\mathbf{S}$, where standardization to zero mean and unit variance is carried out within columns of $\mathbf{S}$. Using absolute values makes minimum- and maximum-values comparable. Matrix $\mathbf{C}$ contains rank-values computed within each column of matrix $\mathbf{S}$. Vector $\mathbf{c}$ contains the most extreme rank value of the j-th row of $\mathbf{C}$ at the j-th position c_j. This can be either the smallest or the largest rank-value. The latter is transformed to a value equal to $N - c_j + 1$. In the following, N_{k-1}corresponds to the number of rows in $\mathbf{S}$ after iteration $(k-1)$, $\boldsymbol{\theta}_k$ is a vector of indices found in iteration k, and N_θ^k is the length of vector $\boldsymbol{\theta}_k$.

The algorithm, initialized with $N_0 = N$ and $k = 1$, can be summarized as follows.

BEGIN
DO WHILE $N_{k-1} - N_\theta^k \geq (1-\alpha)N$

1. Determine indices of the extreme rank-values $\boldsymbol{\theta}_k = I\{\mathbf{c} = min(\mathbf{c})\}$, where I is the indicator function selecting indices of vector elements c_j, $(j \in 1, \ldots, N)$.
2. Remove rows in $\mathbf{S}$ and elements in $\mathbf{c}$ indexed by $\boldsymbol{\theta}_k$.
   IF $N_{k-1} - N_\theta^k = (1-\alpha)N$ THEN DO
   RETURN $\mathbf{S}$
   ELSE DO
   $k = k + 1$

END

3. Compute matrices $\mathbf{D}^\theta$ and $\mathbf{C}^\theta$ which only contain those rows of $\mathbf{D}$ and $\mathbf{C}$ indexed by $\boldsymbol{\theta}_k$.
4. For each row in $\mathbf{C}^\theta$ find index or indices, where $c_l^\theta = min(\mathbf{c})$ and select corresponding element(s) d_l^θ of matrix $\mathbf{D}^\theta$ $(l \in 1, \ldots, n)$.
5. If there are multiple elements d_l^θ use the largest standardized value $max(d_l^\theta)$ for the corresponding row in $\mathbf{D}^\theta$.
6. Among all elements d_l^θ in vector $\mathbf{d}^\theta$ choose the $(1-\alpha)N - (N_{k-1} - N_\theta)$ largest and remove the corresponding rows in matrix $\mathbf{S}$, which then consists of $(1-\alpha)N$ rows.
7. RETURN $\mathbf{S}$

END

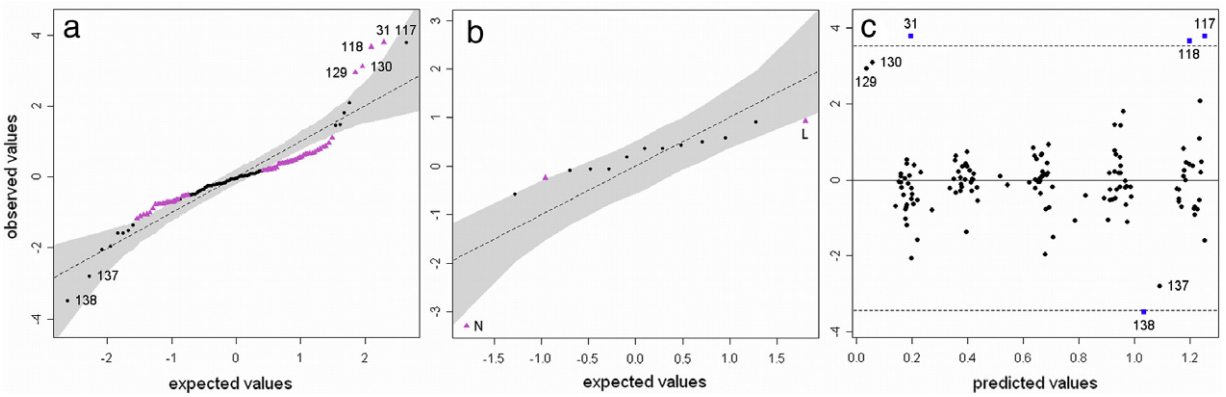

Fig. 2. (a): QQ-plot of studentized conditional residuals for the Cambridge filter data with 95.00% STB; (b): QQ-plot of studentized random laboratory effects for the Cambridge filter data with 95.00% STB; (c): residual plot of studentized conditional residuals vs. predicted values with 95.00% STI.

The quantile-based algorithm (Schützenmeister et al., 2012) as well as the rank-based algorithm can be used to compute the bounds of the second type of simultaneous tolerance bounds: simultaneous tolerance intervals (STI). STIs enclose at least $100(1-\alpha)\%$ of all N simulated residual vectors simultaneously. Computation is much less elaborate compared to STBs, since only the minimum and maximum order statistics (1st and n-th column of $\mathbf{S}$) for each simulated residual vector are used.

Both types of simultaneous tolerance bounds correspond to the null distribution of the respective diagnostic/graphical feature and are helpful when interpreting residual plots. CRs may be illustrated in a QQ-plot, which can be supplemented by an STB. This makes the interpretation of the QQ-plot more objective (Fig. 1e and g, compared to plots above). Using an STB can reveal not only departure from normality, but also outlying observations, i.e. residual points, which are extreme and which lie far outside the $100(1-\alpha)\%$ STB. If such outlying residuals also appear as outliers in residual plots with $100(1-\alpha)\%$ STI (Fig. 1e and f), they are likely to be true outliers. Furthermore, one often observes an increasing residual variance with increasing predicted values $\hat{\mathbf{y}}$. To assess this, Schützenmeister et al. (2012) proposed regressing absolute values or squares of studentized residuals on predicted values. A slope distinctly different from zero would indicate a variance-mean dependence. The regression, denoted as $abs(\tilde{\mathbf{e}}_{obs}) \sim \hat{\mathbf{y}}_{obs}$, can be judged via a $100(1-\alpha)\%$ STB and computed from a $(N \times n)$ matrix $\mathbf{M}$. The j-th row of $\mathbf{M}$ stores the points of the regression line obtained from the regression of the j-th simulated dataset $abs(\tilde{\mathbf{e}}_j) \sim \hat{\mathbf{y}}_j$ $(j = 1, \ldots, N)$. The corresponding STB is computed from $\mathbf{M}$ in the same manner as for QQ-plots. See Fig. 3d for an example.

One particular strength of our method is that it maintains the association between residual points and observations. By using a simulation-based approach to derive the null distribution of residuals, one avoids dependence on asymptotic theory (Longford, 2001). The approach requires reasonably accurate estimates of all variance components, and hence a sufficient number of observed levels for random effects, so it is prudent to study its performance in specific settings by simulation (see Section 6).

5. Examples

5.1. Cambridge filter data

The dataset considered in this section was presented in Rocke (1983), and used by Gumedze et al. (2010) to illustrate the variance shift outlier model (VSOM). It comprises ten samples of Cambridge filter pads with increasing nicotine content. Each of 14 laboratories (lab) analyzed one complete set of filters, i.e. each of the ten nicotine concentrations. The aim of the original study was to assess the reliability of gas chromatography as a first step in analyzing nicotine content (Gumedze et al., 2010). There were 138 measurements, since two values were missing. Gumedze et al. (2010) used the LMM

$$y_{ij} = \mu + \alpha_i + b_j + e_{ij}, \tag{8}$$

where y_{ij} $(i = 1, \ldots, 10;\ j = 1, \ldots, 14)$ is the amount of nicotine in the ij-th sample in milligrams, α_i is the fixed effect of the i-th sample, $b_j \sim N(0, \sigma_b^2)$ is the i.i.d. random effect of the j-th lab, and $e_{ij} \sim N(0, \sigma_e^2)$ is the i.i.d. residual error term for the ij-th measurement. Gumedze et al. identified nine outlying observations (i.e., 9, 31, 109, 117, 118, 129, 130, 137, 138), where four came from lab 14 (N, if letter-coded), which Gumedze et al. identified as an outlier. Christensen et al. (1992), (cited by Gumedze et al. (2010)) identified seven outliers (31, 117, 118, 129, 130, 137, 138) using a case-deletion approach.

Fig. 2 depicts diagnostic plots for the residual analysis of the complete Cambridge filter dataset. From Fig. 2a, one can see that studentized CRs indicate problems, since there are many residuals outside the 95.00% STB. Fig. 2b reveals that lab 14 (N) has by far the largest random lab effect in absolute terms, falling outside the associated 95.00% STB together with two other studentized lab-effects. This plot also reveals that normality of the random lab effects cannot be assumed, because the lab effects do not behave as expected, i.e. they do not scatter around the angle bisecting line (dashed), which indicates their

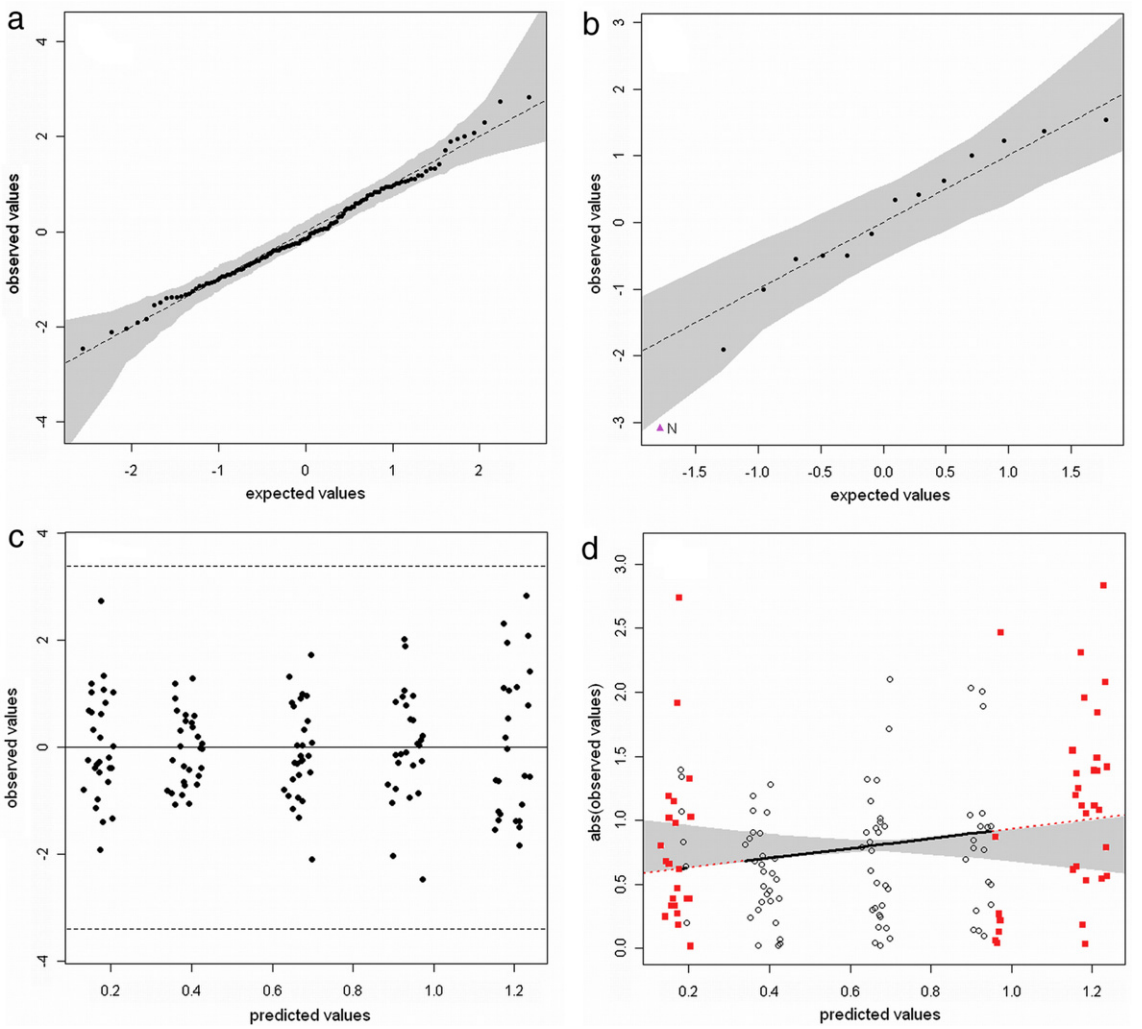

Fig. 3. Plots for the Cambridge filter data, where four variance estimates were used for labs 4 (D), 12 (L), 14 (N), and for the remaining labs. observation 9, 106, 125 were iteratively removed. (a): QQ-plot of studentized conditional residuals (CR) with 95.00% STB; (b): QQ-plot of studentized random lab effect with 95.06% STB; (c): residual plot of studentized CRs with 95.00% STI; (d): plot of absolute values of studentized CRs [$abs(\hat{\mathbf{e}}^*)$] vs. predicted values with 95.02% STB for the observed regression lines. Dotted segments of this regression line exceed the STB.

expected values computed from all N datasets simulated under the null hypothesis. The residual plot of studentized CRs vs. predicted values (Fig. 2c) sheds light on the overall behavior of CRs. There are several residuals located remarkably remote. Three of these residuals are located outside the 95.00% STI for the complete data. Removing the nine outliers identified with the VSOM of Gumedze et al. (2010) did not remedy the problems apparent in the QQ-plot for studentized random effects (not shown). Lab 14 (N) still stands out as an outlier besides other random effects which also exceed the bounds of the STB. Gumedze et al. use a VSOM-term for laboratory 14 to down-weight its influence. Removing the complete data subset from lab 14 made the QQ-plot for the random laboratory-effects conformable with the null distribution of this diagnostic plot (STB) at the cost of removing many informative observations.

A less rigorous way of fitting a model to these data is to estimate lab-specific residual variance parameters. Inspection of these estimates suggests that all labs have the same residual variance, except labs 4 (D), 12 (L) and 14 (N). Thus, we fitted a model with four variance parameters, one for each of the labs D, L, N, and one for the remaining labs. In fact, it is well known that variances may vary among labs, and it is common practice to fit heteroscedastic models to experiments involving several labs (Deutler, 1991; Piepho, 1996). When observation 9, 106, 125 were removed, we obtained diagnostic plots, which we consider acceptable (Fig. 3). In the QQ-plot with 95.00% STB of studentized CRs there is no evidence of non-normality (Fig. 3a). Fig. 3b depicts the QQ-plot of studentized random effects with 95.06% STB. It looks much better than for the original data (Fig. 2b), i.e. random lab effects scatter around the diagonal line. Although lab N still exceeds the STB, the overall appearance meets the assumptions reasonably well.

The plot of studentized CRs vs. predicted values (Fig. 3c) does not reveal outliers but might raise concerns about a possible dependence of the residual variance on predicted values. To further assess whether the residual variance depends on

predicted values, we compute the STB for the regression of studentized CRs on predicted values (see Section 4). Fig. 3d depicts a possible graphical display of the 95.02% STB, useful for assessing whether the residual variance depends on predicted values. The regression line is located slightly outside the associated STB for small and large predicted values (dotted). This violation is borderline, since the regression line for the observed data is located on the bounds of the STB.

We conclude that an LMM with four variance parameters and three observations removed (9, 106, 125) fits the assumptions of the LMM analysis quite well. Thus, less valuable information needs to be discarded compared to the homoscedastic model. Using only four variance parameters in the heteroscedastic model is also more parsimonious than the full heteroscedastic model with 14 variance parameters. In particular, the full heteroscedastic model does not improve the model fit significantly, i.e. the associated likelihood ratio test is neither significant for the complete data ($p = 0.4754$) nor for the data, where observation 9, 106, 125 were removed ($p = 0.4295$). Note that outliers, which were identified by our MC-based graphical approach, do not necessarily have to be deleted. One may simply down-weight these outlying observations by employing extra random effects as proposed by Gumedze et al. (2010) for their VSOM.

5.2. Orthodont data

In this section, we show how STBs can be used to assess normality of single random effects. We additionally show that the type of scaling used for the random terms influences the interpretation of diagnostic plots. The model that we use in this section has random effects that are correlated. We illustrate the simulation approach by applying it to the *Orthodont* data, which comes with the R-package nlme (R Development Core Team, 2011). It is described in Pinheiro and Bates (2000) and comprises the growth records of 27 children (16 male, 11 female) at ages 8–14. The distance between the pituitary and pterygomaxillary fissure was measured every two years from X-ray exposures. We will restrict our interest here to the model *fm2Orth.lme* Pinheiro and Bates (2000, p. 148):

$$y_{ijk} = \alpha_j + \beta_j(x - 11) + a_i + b_i(x - 11) + e_{ijk}, \tag{9}$$

where y_{ijk} is the measured distance of the i-th child, which belongs to the j-th sex, at centered age $k \in \{-3, -1, 1, 3\}$. The parameter α_j is the sex-specific fixed intercept, β_j is the sex-specific fixed slope. Both effects constitute the fixed effect parameter vector $\boldsymbol{\beta} = [\alpha_1, \alpha_2, \beta_1, \beta_2]^T$. Random intercepts a_i and slopes b_j constitute the random effects vector $\mathbf{b} = [a_1, b_1, a_2, b_2, \ldots\ldots, a_{27}, b_{27}]^T$, which is $\mathbf{b} \sim N(0, \mathbf{G})$ distributed independently of the residual errors $\mathbf{e} \sim N(0, \mathbf{R})$, where $\mathbf{R} = \sigma^2\mathbf{I}$. The covariance matrix of random effects a_i and b_i of the i-th child $\boldsymbol{\Sigma}$, making up the blocks of the block-diagonal covariance matrix $\mathbf{G} = \mathbf{I}_{27} \otimes \boldsymbol{\Sigma}$, was estimated as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3.350 & 0.068 \\ 0.068 & 0.033 \end{bmatrix},$$

where the diagonal elements correspond to the variances of the random intercept and random slope, respectively, and the off-diagonal element represents the covariance between both random effects for one subject (child). This covariance was accounted for in the simulation of new data, i.e. the random intercepts and random slopes were drawn from a bivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ using formula (6).

The analysis of the studentized CRs revealed two distinct outlying observations. The 3rd observation of subject M09 (M09.3) and the 1st observation of subject M13 (M13.1) were consecutively identified as outliers. We used a QQ-plot with 95.00% STB ($N = 10,000$ simulations) and the corresponding residual plot of studentized CRs vs. predicted values with 95.00% STI for the identification of both outliers (not shown). To check that confounding of residuals did not lead to misleading conclusions, we also checked the QQ-plot of studentized residuals obtained for the fixed effect analysis of the LMM without observation M09.3 and M13.1. The corresponding 95.07% STB enclosed the observed order statistics completely (not shown). Thus, we concluded that the normality and homoscedasticity assumptions were met for CRs, when observation M09.3 and M13.1 were removed.

Pinheiro and Bates (2000, p. 189) standardized random effects by dividing them by the square root of the respective estimated variance component. They marked individual random effects as outliers, whose absolute values of standardized estimates exceed the 95% quantile of the standard normal distribution. This corresponds to a 10% outlier test. Pinheiro and Bates classified the random intercept effects of subjects F10, F11, and M10 as outlying values. Subject M13 was classified as an outlier for the random slope. Fig. 4a depicts the QQ-plot for the standardized random intercepts, Fig. 4d shows the QQ-plot for the standardized random slopes as presented in Pinheiro and Bates (2000, p. 189). The authors concluded from these plots that the normality assumption is reasonable. Again, it is not clear whether these patterns are consistent with expectation or not. The authors also mention that a few outliers appear to be present, which is not confirmed by our Monte Carlo approach. Fig. 4b depicts the QQ-plot with approximately 90% STBs for the standardized random intercept for each subject. Fig. 4e corresponds to the QQ-plot with approximately 90% STB for the standardized random slopes. We used 90% STBs here and additionally plotted standardized random effects for better comparability to the results presented in Pinheiro and Bates (2000). Fig. 4e reveals the uncertainty inherent in estimating the variance of the random slope, using standardized values. As is apparent, the upper bound for negative values and the lower bound for positive values are practically zero. This results from many slope-variance estimates which are close to zero, obtained in $N = 10,000$ simulations. By using standardized random effects instead of studentized random effects, one does not consider that each random effect estimate

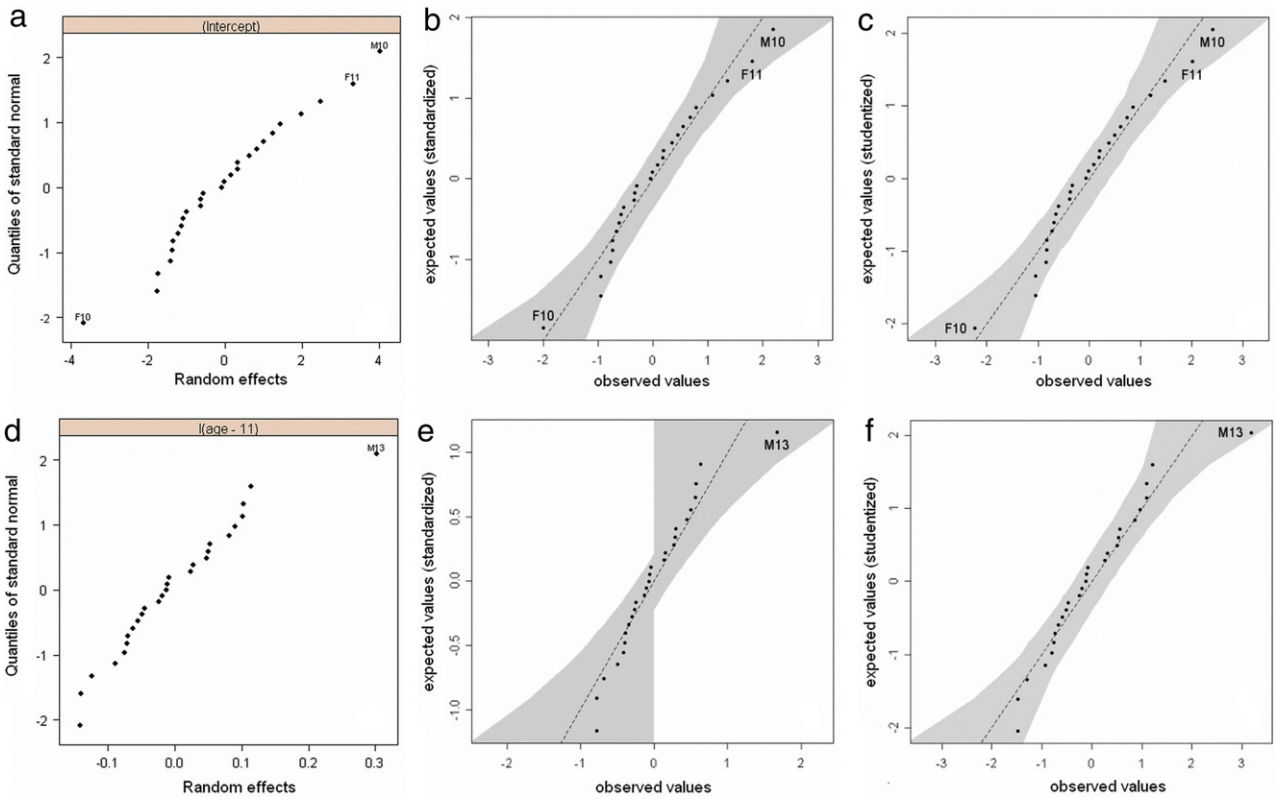

Fig. 4. QQ-plots for random intercept-effects (upper three plots) and random slopes (lower three plots). 1st column (plots a and d): plots of standardized random effects as presented in Pinheiro and Bates (2000, p. 189). 2nd column (plots b and e): QQ-plots of standardized random effects with approximately 90% STB. 3rd column (plots c and f): QQ-plots of studentized random effects with approximately 90% STB.

may have an individual variance. Studentization of random effects, using formula (5), takes this into account (Fig. 4c and f). The overall point pattern is similar to the one obtained from standardization, although there are some scale differences. The associated STB for the random slope is narrower and therefore more meaningful.

Either using standardized values or using studentized values, in both cases the assumption of normality appears to be met and no outlying values seem to be present. There are no values exceeding the bounds of the STBs, which is noteworthy considering the fact that using a value of $\alpha = 0.1$ is a rather liberal choice. It results in a narrower STB, which simultaneously covers approximately $100(1-\alpha)\% = 90\%$ of all $N = 10,000$ vectors of simulated random effects. For this random regression example, it could also be useful to construct an STB for the random regression lines, which combine the information of both random effects (intercept, slope). This is easily achieved by our general method. A possible graphical display is similar to that used in Fig. 3d, although there were 27 regression lines, which would have to be added to the plot.

The QQ-plots with STB (CRs and random effects) for the data, where observations M09.3 and M13.1 were removed, do not reveal any outliers (not shown) and look entirely unsuspicious. The random slope effect of subject M13 is located mid-range, and does not stand out any more.

6. Assessing the empirical size of simultaneous tolerance bounds

We conducted a small simulation study to assess whether our proposed simultaneous tolerance bounds, i.e. STBs and STIs, have empirical error rates (empirical size) conformable with the specified nominal error rate of $\alpha = 5\%$. We consider two different LMMs, both applied to balanced and unbalanced data. The first model (I) is the one used for the toothbrush data (Section 2). More specifically, for the balanced case we investigated three designs, comprising $g = 32$, 16 or 8 subjects, each having $r = 4$, 8 or 16 repeated measures, respectively, using Eq. (1). The 2nd model (II) was fitted to three designs with g groups and r replicates per group, where $r \times g = 36$, and can be written as $y_{ij} = \mu + a_i + e_{ik}$, where observation y_{ij} corresponds to the j-th replicate of the i-th group, μ is the fixed intercept, a_i is the i.i.d. $N(0, \sigma_a^2)$ distributed random group-effect, e_{ik} is the i.i.d. $N(0, \sigma_e^2)$ distributed ij-th error term. We simulated random effects with variance $\sigma_a^2 = \gamma$ and residual errors with variance $\sigma_e^2 = 1$ from normal distributions with expected values equal to zero. Since the random part of the LMM is independent of the fixed effects (see Section 4), each dataset was simulated according to Eq. (6). Table 1 contains the results of the simulation study for the balanced case.

Table 1
Empirical error rates (type I errors $\alpha = 0.05$) for balanced designs regarding approximately 95% simultaneous tolerance band (STB) of studentized conditional residuals (CR), 95% simultaneous tolerance interval (STI) of studentized CRs, and approximately 95% STB for studentized random effects (RE). Model I corresponds to the toothbrush dataset with g subjects and r measurements per subject, Model II corresponds to a one-way random effects ANOVA with g groups and r replicates per group.

Model	g	r	Balanced designs: $\gamma = \sigma_a^2/\sigma_e^2$								
			$\gamma = 0.1$			$\gamma = 1$			$\gamma = 10$		
			STB CR	STI CR	STB RE	STB CR	STI CR	STB RE	STB CR	STI CR	STB RE
	32	4	0.066	0.051	0.042	0.048	0.069	0.054	0.058	0.054	0.055
I	16	8	0.057	0.058	0.037	0.058	0.037	0.057	0.050	0.046	0.060
	8	16	0.059	0.051	0.034	0.069	0.042	0.049	0.055	0.064	0.053
	12	3	0.050	0.047	0.039	0.054	0.051	0.039	0.047	0.049	0.048
II	6	6	0.067	0.053	0.043	0.054	0.052	0.051	0.052	0.043	0.058
	3	12	0.061	0.044	0.028	0.057	0.050	0.030	0.053	0.046	0.047

Table 2
Empirical error rates (type I errors $\alpha = 0.05$) for approximately 95% simultaneous tolerance band (STB) of studentized conditional residuals (CR), 95% simultaneous tolerance interval (STI) of studentized CRs, and approximately 95% STB for studentized random effects (RE). Model I corresponds to the toothbrush dataset with g subjects and r measurements per subject, Model II corresponds to a one-way random effects ANOVA with g groups and r replicates per group. Model I and model II were applied to unbalanced data, i.e. group sizes were unequal.

Model	g	r	Unbalanced designs: $\gamma = \sigma_a^2/\sigma_e^2$								
			$\gamma = 0.1$			$\gamma = 1$			$\gamma = 10$		
			STB CR	STI CR	STB RE	STB CR	STI CR	STB RE	STB CR	STI CR	STB RE
	32	2–9	0.057	0.041	0.061	0.053	0.056	0.056	0.065	0.044	0.061
I	16	3–16	0.059	0.047	0.041	0.060	0.046	0.056	0.062	0.053	0.052
	8	6–23	0.075	0.064	0.050	0.060	0.052	0.052	0.054	0.058	0.039
	12	1–6	0.065	0.052	0.052	0.051	0.048	0.056	0.062	0.039	0.033
II	6	2–11	0.050	0.056	0.044	0.045	0.051	0.054	0.047	0.042	0.053
	3	6–18	0.058	0.046	0.051	0.052	0.044	0.021	0.048	0.044	0.031

For the unbalanced case, we used the same number of groups as for the balanced data, but varied the number of repeated measures among the g groups. The simulated data under the null hypothesis was created the same way as the balanced data, i.e. random effects a_i and e_{ij} were drawn from normal distributions $N(0, \sigma_a^2)$ and $N(0, \sigma_e^2)$, respectively. For model I, we introduced unbalancedness by creating group sizes ranging from $r = 2$ to $r = 9$ for the $g = 32$ data, group sizes from $r = 3$ to $r = 16$ for the $g = 16$ data, and group sizes from $r = 6$ to $r = 23$ for the $g = 8$ data. Model II was created the same way as model I. Here we obtained unbalanced data sets by creating group sizes from $r = 1$ to $r = 6$ for the $g = 12$ data, group sizes from $r = 2$ to $r = 11$ for the $g = 6$ data, and group sizes $r = 6, r = 12, r = 18$ for the $g = 3$ data. Table 2 contains the results of the simulation study for the unbalanced datasets.

We generated 1000 datasets for each combination of experimental design and ratio of variance components γ, and constructed the approximately 95% STBs for studentized CRs and studentized random effects, and the 95% STI for studentized CRs from 5000 (inner) simulations. Whenever at least one residual fell outside of these simultaneous tolerance bounds, we classified it as non-conformable with the null hypothesis. We expected that approximately 5% of all 1000 outer simulations would reveal such violations of the simultaneous tolerance bounds. A 95% tolerance interval for the empirical size (error rate under the null hypothesis) can be constructed from the binomial distribution, since the STIs and STBs classify a residual vector as either acceptable (all points enclosed) or not (at least one point outside), which can be regarded a Bernoulli experiment. For $\alpha = 5\%$ and $n = 1000$, the associated 95% tolerance interval is equal to $[0.0365;\ 0.0635]$.

This simulation study revealed no systematic deviations from the expected error rates, neither for the balanced (Table 1) nor the unbalanced (Table 2) designs. There are few empirical error rates, which are not enclosed by the associated 95% tolerance interval. These values primarily occur for experimental designs where only a few groups are present. van Eeuwijk (1995) stressed that distributional properties of random effects, where less than ten degrees of freedom are available for estimating the associated variance component, cannot be properly checked. Those values outside of the 95% tolerance interval occurred predominantly for designs, for which less than ten degrees of freedom were available for prediction of the pertinent variance component.

7. Discussion

In this paper, we present a Monte Carlo approach to residual analysis for LMMs, which is widely applicable. We see the main advantage of this approach in adding some objectivity to the interpretation of different diagnostic plots for various types of LMM residuals. Informal/diagnostic procedures like QQ-plots or plots of residuals vs. predicted values, are frequently used to assess normality or homoscedasticity. With such plots, a data analyst is generally faced with the problem of judging an observed pattern to be indicative of violating either normality or homoscedasticity or not. Simultaneous tolerance bounds (STB, STI) allow even less experienced users to assess these plots more objectively.

Different types of LMM residuals are not independent of each other. To rule out the possibility that confounding has an influence on the conclusions drawn from e.g. QQ-plots with STB for assessing normality, one can apply a fixed-effect analysis of the LMM, where random effects are specified as fixed effects. An alternative procedure, which also accounts for a possible confounding of LMM residuals, is to use linear transformations as suggested by Hilden-Minton (1995) and Nobre and Singer (2007). One major drawback of this approach is the loss of interpretability, since transformed residuals do not refer to single observations, and hence possible outliers may be missed. Moreover, power is lost because the linear transformation of n observations to $(n - r)$ least-/unconfounded residuals ($r = \text{rank}\,[\mathbf{X}|\mathbf{Z}]$) reduces the number of observations and introduces further *"supernormality"*. There is no confounding in the fixed-effects analysis, residuals refer to observations, and it is computationally less expensive than the LMM analysis.

Our approach can be used for all types of LMM residuals. Besides CRs, it also applies to estimates of individual random effects **b** or **Zb**. The former can be used to identify outliers for a specific random effect, whereas the latter is useful for assessing the combined behavior of multiple random effects. In Section 5.1, we showed that inspecting QQ-plots of studentized random effects may help in specifying a proper covariance structure **R** of the error terms, which better reflects model assumptions. The normality of random effects and errors is a prerequisite of any LMM analysis, which makes checking their normality mandatory. In Section 5.2, we demonstrated that using studentized random effects should be preferred over standardized values, since the latter ignores the possibility that different estimated/predicted random effects may have different variances, which need to be accounted for and hence is more likely to lead to wrong conclusions about individual random effects.

Our approach also provides a means for assessing the homoscedasticity of CRs. We propose using an STI, which can be added to plots of CRs vs. predicted values. Possible outliers can be objectively identified using such plots, and a possible dependence of the residual variance on predicted values can also be inferred. The latter may be assessed by an STB for the regression of absolute values (or squares) of CRs on predicted values as shown in Section 5.1.

Besides checking model assumptions, our method also provides a means for detecting possible outliers. Outliers may occur on the level of single observations or on the level of random effects, e.g. in Section 5.1 laboratory 14 (N) had by far the largest random laboratory effect in terms of absolute values. Since our approach can be used to construct various diagnostic plots, one can and should combine the information inferred from different types of diagnostic plots. We would like to stress that a suspiciously deviant CR is more likely to be truly outlying, if it appears simultaneously as an outlier in the residual plot (outside the STI), and in the QQ-plot (outside the STB). Outliers identified this way can be either removed or down-weighted. The latter could be done in the same manner as presented in Gumedze et al. (2010) by introducing random effects for each of the outlying observations or for a group of observations. This approach avoids losing valuable information while introducing additional parameters (variance components) and the decision whether to remove or down-weight can be based on goodness-of-fit criteria, e.g. AIC or BIC.

Our approach is applicable to a large variety of LMMs as the example analyses show. It can be applied to models with non-zero covariances among random effects (*Orthodont* data), to models with heterogeneous variances (*Cambridge filter* data), even to models with autocorrelated error structure, e.g. time series models or models where errors are correlated in space (geostatistics). The key is to integrate all model assumptions in the simulation of the data under the null hypothesis (parametric bootstrap), which can be done by applying Eqs. (6) or (7). The conceptual simplicity of our approach makes it even feasible for generalized linear (mixed) models (GLM, GLMM). But with GLM/GLMMs it becomes necessary to also simulate fixed effects of the linear predictor, as the variance depends functionally on the value of the linear predictor. Also, convergence problems are more common with these types of models, particularly with sparse count data.

We also proposed an alternative algorithm for computing simultaneous tolerance bounds. The quantile-based algorithm (Schützenmeister et al., 2012) is somewhat time consuming, particularly for LMMs, where several simultaneous tolerance bounds may need to be computed. The more simulation cycles there are, the more elaborate is the computation of point-wise tolerance intervals needed to compute the simultaneous coverage of all simulated vectors. The rank-based algorithm described in Section 4 produces simultaneous tolerance bounds, which are practically indistinguishable from the bounds computed with the quantile-based algorithm if N gets large, where large means $N \geq 20,000$. For datasets with fewer observations, N may be smaller than 20,000, which we inferred from simulation of various datasets with differing sizes. For the examples presented in this paper, we never used more than $N = 10,000$ simulations when the quantile-based algorithm (Schützenmeister et al., 2012) was used to compute simultaneous tolerance bounds.

Simulation approaches are time consuming, especially for complex LMMs. Fortunately, such approaches are tailored for parallel-computing. The time spent on the simulation can be reduced by factor of K, which corresponds to the number of threads used for the application. Therefore, we think that simulation approaches, like the one presented here, will gain even more importance in the foreseeable future. We found our approach extremely helpful when processing high-dimensional datasets from metabolic profiling (Römisch-Margl et al., 2010) and for which each variable is analyzed separately. Diagnostic plots can be assessed rather intuitively and provide an excellent tool for the residual analysis of LMMs.

Acknowledgments

We thank Juvencio Santos Nobre and Julio da Motta Singer for the permission to use the toothbrush dataset and for supplying their R-code which proved very helpful in implementing our procedure. Furthermore, we would like to thank

two anonymous reviewers for their suggestions and comments, which helped improve an earlier version of this paper. We also thank Joseph Ogutu for valuable comments and corrections.

References

Atkinson, A.C., 1981. Two graphical displays for outlying and influential observations in regression. Biometrika 68, 13–20.
Atkinson, A.C., 1985. Plots, Transformations and Regression. Oxford University Press, Oxford.
Christensen, R., Pearson, L.M., Johnson, W., 1992. Case-deletion diagnostics for mixed models. Technometrics 34, 38–45.
Cochran, W.G., Cox, D.R., 1957. Experimental Designs, second ed.. Wiley, New York.
Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman and Hall, London.
Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.
Deutler, T., 1991. Grubbs-type estimators for reproducibility variances in an interlaboratory test study. Journal of Quality Technology 23, 324–333.
Dufour, J.-M., Farhat, A., Gardiol, L., Khalaf, L., 1998. Simulation-based finite sample normality tests in linear regression. Econometrics Journal 1, 154–173.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.
Gumedze, F.N., Welham, S.J., Gogel, B.J., Thompson, R., 2010. A variance shift model for detection of outliers in the linear mixed model. Computational Statistics and Data Analysis 54, 2128–2144.
Haslett, J., Dillane, D., 2004. Application of delete=replace to deletion diagnostics for variance component estimation in the linear mixed model. Journal of the Royal Statistical Society, Series B 66, 131–143.
Henderson, C.R., 1950. Estimation of genetic parameters Abstract. Ann.Math.Sci. 21, 309–310.
Huang, X., 2011. Detecting random-effects model misspecification via coarsened data. Computational Statistics and Data Aanalysis 55, 703–714.
Hilden-Minton, J.A., 1995. Multilevel diagnostics for mixed and hierarchical linear models. PhD Thesis. University of California, Los Angeles.
Kackar, A.N., Harville, D.A., 1984. Unbiasedness of two-stage estimation and precision procedures for mixed linear models. J. Amer. Statist. Assoc. 10, 1249–1261.
Laird, N.M., Ware, J.H., 1982. Random effects models for longitudinal data. Biometrics 38, 963–974.
Lange, N., Ryan, L., 1989. Assessing normality in random effects models. Ann. Statist. 17, 624–642.
Longford, N.T., 2001. Simulation-based diagnostics in random-coefficient models. Journal of the Royal Statistical Society, Series A 164, 259–273.
Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.
Nobre, J.S., Singer, J.M., 2007. Residual analysis for linear mixed models. Biometrical Journal 49, 863–875.
Piepho, H.P., 1996. Weighted estimates of interlaboratory consensus values. Computational Statistics and Data Analysis 22, 471–479.
Piepho, H.P., 2009. Ridge regression and extensions for genome-wide selection in maize. Crop Science 49, 1165–1176.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in *S* and S-PLUS. Springer, New York.
R Development Core Team 2011. R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
Robinson, G.K., 1991. That BLUP is a good thing: The estimation of random effects. Statistical Science 6, 15–51.
Rocke, D.M., 1983. Robust statistical analysis of interlaboratory studies. Biometrika 70, 421–431.
Römisch-Margl, L., Spielbauer, G., Schützenmeister, A., Schwab, W., Piepho, H.P., Genschel, U., Gierl, A., 2010. Heterotic patterns of sugar and amino acid components in developing maize kernels. Theoretical and Applied Genetics 120, 369–381.
Schützenmeister, A., Jensen, U., Piepho, H.P., 2012. Checking the assumptions of normality and homoscedasticity in the general linear model using diagnostic plots. Communications in Statistics–Simulation and Computation 41, 141–154.
Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. John Wiley & Sons, New York.
Seber, G.A.F., 1977. Linear Regression Analysis. John Wiley & Sons, New York.
Shi, L., Huang, M., 2011. Stepwise local influence analysis. Computational Statistics and Data Analysis 55, 973–982.
Shi, L., Chen, H., 2012. Deletion, replacement and mean-shift for diagnostics in linear mixed models. Computational Statistics and Data Analysis 56, 202–208.
Singer, J.M., Andrade, D.F., 1997. Regression models for the analysis of pretest/posttest data. Biometrics 53, 729–735.
Singer, J.M., Nobre, J.S., Sef, H.C., 2004. Regression models for pretest/posttest data in blocks. Statistical Modelling 4, 324–338.
Thompson, R., 1985. A note on restricted maximum likelihood estimation with an alternative outlier model. Journal of the Royal Statistical Society, Series B 47, 53–55.
van Eeuwijk, F.A., 1995. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. Euphytica 84, 1–7.
Verbeke, G., Lesaffre, E., 1996. A linear mixed-effects model with heterogeneity in the random-effects population. J. Amer. Statist. Assoc. 91, 217–221.
Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, Berlin.