



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.

© 2011 Elsevier B.V. All rights reserved.

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).

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

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

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

Download English Version:

<https://daneshyari.com/en/article/416436>

Download Persian Version:

<https://daneshyari.com/article/416436>

[Daneshyari.com](https://daneshyari.com)