R function for residual analysis in linear mixed models: lmmresid

Juvêncio S. Nobre[1,‡] and **Julio M. Singer**[2,§]

[1]Departamento de Estatística e Matemática Aplicada,
Universidade Federal do Ceará, Fortaleza, Brazil

[2]Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil

[‡]`juvencio@ufc.br` [§]`jmsinger@ime.usp.br`

1 Introduction

Our objective is to illustrate the use of a function written in the `R` language (R Development Core Team, 2009) for residual analysis in linear mixed models as presented in Nobre and Singer (2007). To use the routines it is necessary install the basic `R` software and the packages `Matrix`, `lattice` and `lme4` that may be obtained from

`www.r-project.org/`.

The function may be obtained from `www.dema.ufc.br/~juvencio`.

The model we are interested in may be expressed as

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

where $\mathbf{y}_i$ is a $m_i \times 1$ vector of observations (*response profile*) for the i-th unit, $i = 1, \dots, n$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is a $p \times 1$ vector of unknown population parameters (*fixed effects*), $\mathbf{X}_i$ is a $m_i \times p$ known specification matrix corresponding to the fixed effects, $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})^\top$ is a $q \times 1$ vector of unknown random parameters (*random effects*), $\mathbf{Z}_i$ is a $m_i \times q$ known specification matrix corresponding to the random effects and $\mathbf{e}_i$ is an $m_i \times 1$ vector of random errors. Usually one assumes that

$$\mathbf{b}_1, ..., \mathbf{b}_n \overset{\text{iid}}{\sim} \mathcal{N}_q(\mathbf{0}, \sigma^2\mathbf{G}) \;\text{ e }\; \mathbf{e}_i \overset{\text{ind}}{\sim} \mathcal{N}_{m_i}(\mathbf{0}, \sigma^2\mathbf{R}_i), \quad i = 1, ..., n, \tag{2}$$

with $\mathbf{b}_i$ and $\mathbf{e}_i$ independent, $\mathbf{G}$ and $\mathbf{R}_i$ being $(q \times q)$ and $(m_i \times m_i)$ positive definite matrices respectively, with elements expressed as functions of a vector of covariance parameters, $\boldsymbol{\theta}$, not functionally related to $\boldsymbol{\beta}$. Letting $\mathbf{y} = (\mathbf{y}_1^\top, \cdots, \mathbf{y}_n^\top)^\top$, $\mathbf{X} = (\mathbf{X}_1^\top, \cdots, \mathbf{X}_n^\top)^\top$, $\mathbf{Z} = \bigoplus_{i=1}^n \mathbf{Z}_i$, where $\bigoplus$ represents the direct sum, $\mathbf{b} = (\mathbf{b}_1^\top, \cdots, \mathbf{b}_n^\top)^\top$ and $\mathbf{e} = (\mathbf{e}_1^\top, \cdots, \mathbf{e}_n^\top)^\top$, we can write model (1) more compactly as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e}. \tag{3}$$

This implies that

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{e} \end{bmatrix} \sim \mathcal{N}_{M+N}\left(\begin{bmatrix} \mathbf{0}_M \\ \mathbf{0}_N \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{D} & \mathbf{0}_{M\times N} \\ \mathbf{0}_{n\times M} & \sigma^2\boldsymbol{\Sigma} \end{bmatrix} \right),$$

where $N = \sum_{i=1}^n m_i$, $M = nq$, $\mathbf{D} = \mathbf{I}_n \bigotimes \mathbf{G}$ and $\boldsymbol{\Sigma} = \bigoplus_{i=1}^n \mathbf{R}_i$, with $\bigotimes$ denoting the Kronecker product and $\mathbf{I}_n$, , the identity matrix of order n.

Given the model specification and assuming that the covariance matrix $\boldsymbol{\Omega}(\boldsymbol{\theta}) = \sigma^2(\mathbf{Z}\mathbf{D}\mathbf{Z}^\top + \boldsymbol{\Sigma})$, with $\mathbf{D}$ and $\boldsymbol{\Sigma}$ known, then the best linear unbiased estimators (BLUE) of the fixed effects parameters $\boldsymbol{\beta}$ and best linear predictors (BLUP) of the random effects $\mathbf{b}_i$ may be obtained as the solutions to Henderson's equations

$$\begin{pmatrix} \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} \end{pmatrix}, \tag{4}$$

namely,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y} \\ \widehat{\mathbf{b}} &= \mathbf{D}\mathbf{Z}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \end{aligned}$$

In practice the covariance matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}$ are unknown. Empirical BLUE or BLUP may be obtained by using consistent estimates of $\mathbf{D}$ and $\boldsymbol{\Sigma}$.

2 Residual analysis for linear mixed models

Residuals are frequently used to evaluate the validity of the assumption of models. For example, in normal linear models residuals are used to verify linearity of effects, normality, independence, homoskedasticity of the errors and presence of outliers or influent observations. Since mixed models have two sources of variation ($\mathbf{e}$ and $\mathbf{b}$), different types of residuals may be defined and the corresponding analysis is more complex. In linear mixed models, there are three types of residuals, namely

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

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

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

According to Hilden-Minton (1995) a residual is said to be confounded for a specific type of error if it also depends on errors different from those that it is supposed to predict. In linear mixed models like (3), conditional residuals and the BLUP are confounded (Nobre and Singer, 2007). This implies, for example, that $\widehat{\mathbf{e}}$ may not be adequate to check for normality of $\mathbf{e}$ since when $\mathbf{b}$ is grossly non-normal, $\widehat{\mathbf{e}}$ may not present a normal beahavior even when $\mathbf{e}$ is normal (Nobre and Singer, 2007, Section 4). Following the suggestion of Hilden-Minton (1995) we consider conditional least confounded residuals, obtained as linear combinations of the conditional residuals that minimize the proportion of their variance due to the random effects. For details, see Nobre and Singer (2007).

Each type of residual is useful to evaluate some assumption of model (1), as indicated in Table 1, where $\widehat{\mathbf{R}}_i = \widehat{\mathbf{V}}_i^{-1/2}\widehat{\boldsymbol{\xi}}_i$, $\widehat{\boldsymbol{\xi}}_i = \mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}$, $\hat{e}_k^*$ represents the k-th element of the vector of standardized conditional residuals and $|\mathrm{d}_{\max}|$ represents the normalized eigenvector associated with the direction of largest normal curvature of the influence graph under the perturbation of the covariance matrix of the random effects (see Nobre and Singer, 2007, for details).

Table 1: Uses of residuals for diagnostic purposes

Diagnostic for	Type of residual	Plot
Linearity of effects ($\mathbb{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$)	Marginal	$\widehat{\boldsymbol{\xi}}_k$ vs. explanatory variables
Within-subjects covariance matrix ($\mathbf{V}_i$)	Marginal	$\lvert\lvert\mathbf{I}_{n_i} - \widehat{\mathbf{R}}_i\widehat{\mathbf{R}}_i^\top\rvert\rvert^2$ vs. subject indices
Presence of outlying observations	Conditional	$\hat{e}_k^*$ vs. observation indices
Homoskedasticity of conditional errors ($\mathbf{e}_i$)	Conditional	$\hat{e}_k^*$ vs. fitted values
Normality of conditional errors ($\mathbf{e}_i$)	Conditional	QQ for least confounded residuals
Presence of outlying subjects	EBLUP	$\widehat{\mathbf{b}}_i\widehat{\mathbb{V}}[\widehat{\mathbf{b}}_i - \mathbf{b}_i]\hat{\mathbf{b}}_i$ vs. subject indices
Random effects covariance structure ($\mathbf{G}$)	EBLUP	$\lvert\mathrm{d}_{\max}\rvert$ vs. subject indices
Normality of the random effects ($\mathbf{b}_i$)	EBLUP	Weighted QQ for $\widehat{\mathbf{b}}_i$

3 Use of the lmmresid function

i) The data should be organized according to the format in Table 2 and loaded into `R`.

ii) The package `lme4` should be loaded via the command `require(lme4)`.

iii) The model of interest should be fitted via the `lme4` package and placed in an object (`fit`, for example).

iv) A variable with the labels of subjects should be created (`subject`, for example).

v) The residual plots are then obtained via the function `lmmresid` with `fit` and `subject` as arguments.

Table 2: Data organization

`Variable1`	`Variable2`	...	`Variablep`
a_{11}	a_{21}	...	a_{p1}
⋮	⋮	⋮	...
a_{1n}	a_{2n}	...	a_{pn}

As an illustration, consider the data set in Nobre and Singer (2007):

```
data<-read.csv2("http://www.dema.ufc.br/~juvencio/dadost.csv",dec=”,”,header=T)
```

The data set is organized as follows

To reproduce the plots in Nobre and Singer (2007), use the following commands

```
require(lme4)
attach(data)
toot<-as.factor(Tootbrush)
session<-as.factor(Session)
subject<-as.factor(subject)
data<-data.frame(data)
fit<-lmer(log(y)~log(x)+toot-1+(1|subject),method="REML")
lmmresid(fit,subject).
```

Toothbrush	Session	x	y	Subject
Conventional	1st	1.05	1.00	1
Conventional	2nd	1.13	0.84	1
Conventional	3rd	1.15	0.86	1
Conventional	4th	1.13	0.94	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Monoblock	1st	1.15	1.00	32
Monoblock	2nd	1.23	1.11	32
Monoblock	3rd	1.15	1.07	32
Monoblock	4th	1.26	1.00	32

Acknowledgements

This research received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Edital Universal 2007 # 479164/2007-0) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil.

References

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

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

R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.