Recent Developments in HLMdiag

Adam Loy
Iowa State University

Abstract

The **HLMdiag** package provides diagnostic measures for linear mixed and hierarchical models fit using `lmer` in the **lme4** package. It take inspiration from `influence.measures` for regression models fit using `lm` for an intuitive presentation of deletion diagnostics. The first versions of **HLMdiag** served to develop the framework around which the package is built, but suffered from slow performance. Using C++, deletion diagnostics have since been implemented affordably.

Keywords: **HLMdiag**, deletion diagnostics.

1. Introduction

The **HLMdiag** package (Loy 2012) provides a framework to obtain residuals and deletion diagnostics from two-level linear mixed and hierarchical models in R (R Core Team 2012) fit by `lmer` in the **lme4** package (Bates, Maechler, and Bolker 2011). In the early versions, deletion diagnostics were not 'usable' as they were not computationally affordable. This due to naive implementation, based on iteratively refitting a model based on a reduced data set. New versions ($> 0.1.5$) implement deletion diagnostics based on the building blocks developed by Christensen, Pearson, and Johnson (1992), Zewotir (2008), and Haslett and Dillane (2004). Additional computational speed has been achieved through the use of sparse matrices using the **Matrix** package (Bates and Maechler 2012), combining R and C++ using the **Rcpp** package (Eddelbuettel and François 2011), and the Armadillo C++ library (Sanderson 2010) via **RcppArmadillo** (Francois, Eddelbuettel, and Bates 2012).

In this paper we provide an overview of the new developments in **HLMdiag**. In Section 2 include a brief overview of the models and detail methodology used to increase computational speed. Section 3 and 4 discuss changes to **HLMdiag**, providing an example, and provides timings comparisons for the implementations. Section 5 concludes this paper.

2. Methods

2.1. The model

We consider a two-level linear mixed model with nested error structure of the form

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times q)}{\boldsymbol{Z}}\underset{(q\times 1)}{\boldsymbol{b}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{1}$$

where $\boldsymbol{b} \sim \mathcal{N}(0,\ \sigma^2 \boldsymbol{D})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\ \sigma^2 \boldsymbol{I})$. Notice that in this formulation, $\boldsymbol{D}$ is a matrix comprised of the ratio of variance components, which we denote using the vector $\boldsymbol{\theta}$. We additionally assume that $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$ are mutually independent. This model specifications results in the following marginal distribution of $\boldsymbol{y}$

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X\beta},\ \sigma^2 \boldsymbol{V}) \tag{2}$$

where $\boldsymbol{V} = \boldsymbol{I} + \boldsymbol{ZDZ}'$. This model specification matches that of `lmer`.

2.2. Overview of deletion diagnostics

Deletion diagnostics have become a common way to assess the sensitivity of the model to the observed data. These diagnostics summarize how the model changes based on the exclusion of one or more observations. In the mixed/hierarchical model setting, it is recommended to run deletion both at the individual (observation) and group (sampling unit) levels. **HLMdiag** includes on deletion diagnostics focusing on specific aspects of the model

- estimates of the fitted values – Cook's distance and multivariate DFFITS (MDFFITS)

$$\mathrm{C}_I\left(\widehat{\boldsymbol{\beta}}\right) = \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[I]}\right)' \boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[I]}\right) / \left(p\widehat{\sigma}^2\right) \tag{3}$$

$$\mathrm{MDFFITS}_I\left(\widehat{\boldsymbol{\beta}}\right) = \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[I]}\right)' \boldsymbol{X}'_{[I]}\widehat{\boldsymbol{V}}^{-1}_{[I]}\boldsymbol{X}_{[I]} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[I]}\right) / \left(p\widehat{\sigma}^2_{[I]}\right) \tag{4}$$

- precision of estimates of the fitted values – covariance trace (COVTRACE) and covariance ratio (COVRATIO)

$$\mathrm{COVTRACE}_I\left(\widehat{\boldsymbol{\beta}}\right) = \frac{\widehat{\sigma}^2_{[I]}}{\widehat{\sigma}^2} \cdot \det\left(\left(\boldsymbol{X}'_{[I]}\widehat{\boldsymbol{V}}^{-1}_{[I]}\boldsymbol{X}_{[I]}\right)^{-1}\right) \det\left(\left(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\right)^{-1} \tag{5}$$

$$\mathrm{COVRATIO}_I\left(\widehat{\boldsymbol{\beta}}\right) = \frac{\widehat{\sigma}^2_{[I]}}{\widehat{\sigma}^2} \cdot \left|\mathrm{trace}\left(\left(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)\left(\boldsymbol{X}'_{[I]}\widehat{\boldsymbol{V}}^{-1}_{[I]}\boldsymbol{X}_{[I]}\right)^{-1}\right) - p\right| \tag{6}$$

- predictions of the random effects – Cook's distance

$$\mathrm{C}_I\left(\widehat{\boldsymbol{b}}\right) = \left(\widehat{\boldsymbol{b}} - \widehat{\boldsymbol{b}}_{[I]}\right)' \boldsymbol{D}^{-1} \left(\widehat{\boldsymbol{b}} - \widehat{\boldsymbol{b}}_{[I]}\right) / \widehat{\sigma}^2 \tag{7}$$

- estimates of the variance components – not yet implemented efficiently

Christensen *et al.* (1992) extended Cook's distance and COVTRACE to the fixed effects and variance components of a mixed model, developing basic building blocks for more efficient computation for single case deletion. Banerjee and Frees (1997) extended Cook's distance using partial influence to account for the deletion of a group of observations, developing simlar building blocks. Finally, Zewotir and Galpin (2005) refined these building blocks and Zewotir (2008) generalized the building blocks for both single case and group deletion. It is from this work that the building blocks used in **HLMdiag** are taken.

2.3. Building blocks of deletion diagnostics

In **HLMdiag**, deletion diagnostics are calculated in two steps. The first step extracts and calculates the building blocks of the requested deletion diagnostics and is carried out using the

Aspect	**Building Blocks**
estimates of the fitted values	$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{[I]}$, $\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}$ or $\boldsymbol{X}'_{[I]}\widehat{\boldsymbol{V}}_{[I]}^{-1}\boldsymbol{X}_{[I]}$, and $\widehat{\sigma}^2$ or $\widehat{\sigma}^2_{[I]}$
precision of estimate of fitted values	$\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}$, $\boldsymbol{X}'_{[I]}\widehat{\boldsymbol{V}}_{[I]}^{-1}\boldsymbol{X}_{[I]}$, $\widehat{\sigma}^2$, and $\widehat{\sigma}^2_{[I]}$
predicted random effects	$\widehat{\boldsymbol{b}} - \widehat{\boldsymbol{b}}_{[I]}$, $\boldsymbol{D}^{-1}$

Table 1: The building blocks of deletion diagnostics for mixed/hierarchical linear models.

`case_delete` function. The second step is to calculate the requested diagnostic(s) and is carried out using either `diagnostics` or the individual diagnostics functions (e.g., `cooksd_hlm`). Affordable computation of diagnostics using this two-step approach requires building blocks that can be extracted/calculated from the original model fit, eliminating the need for a full refit. The building blocks of deletion diagnostics for mixed linear models can be summarized in the table 1.

2.4. Computational building blocks

It is clear that the building blocks presented in table 1 are not computational formulas, but looking to the literature we can adopt affordable formulas for these building blocks. Adopting the notation of Zewotir (2008) we use the subscript $[I]$ to the deletion of elements in the rows and columns indexed by I, and (I) to denote a matrix or vector with the rows indexed by I removed.

Affordable computational building blocks for diagnostics relating to the fixed effects rely on rewriting

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \boldsymbol{V}_{II} & \boldsymbol{V}'_{I(I)} \\ \boldsymbol{V}_{I(I)} & \boldsymbol{V}_{[I]} \end{bmatrix}$$

as a partioned matrix, where $\boldsymbol{V}_{[I]}$ denotes the matrix $\boldsymbol{V}$ with the elements indexed by I removed, $\boldsymbol{V}_{II}$ are the deleted elements, and $V_{I[I]}$ is comprised of the elements from the columns of $\boldsymbol{V}$ from which elements were deleted. Additionally, the results of Christensen *et al.* (1992) and Zewotir (2008) give

Building block 1 $\boldsymbol{X}'_{(I)}\widehat{\boldsymbol{V}}_{[I]}^{-1}\boldsymbol{X}_{(I)} = \boldsymbol{X}'V^{-1}\boldsymbol{X} - \boldsymbol{X}'\boldsymbol{V}_{I(I)}\boldsymbol{V}_{[II]}^{-1}\boldsymbol{V}'_{I(I)}\boldsymbol{X}$

To obtain an expression for the change in fixed effects estimates, we must first define the matrix $\boldsymbol{P} = \boldsymbol{V}^{-1}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})$

References

Banerjee M, Frees E (1997). "Influence Diagnostics for Linear Longitudinal Models." *Journal of the American Statistical Association*, **92**(439), 999–1005.

Bates D, Maechler M (2012). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.0-6, URL http://CRAN.R-project.org/package=Matrix.

Bates D, Maechler M, Bolker B (2011). ***lme4****: Linear mixed-effects models using S4 classes.* R package version 0.999375-42, URL http://CRAN.R-project.org/package=lme4.

Christensen R, Pearson L, Johnson W (1992). "Case-deletion diagnostics for mixed models." *Technometrics*, **34**(1), 38–45.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. URL http://www.jstatsoft.org/v40/i08/.

Francois R, Eddelbuettel D, Bates D (2012). ***RcppArmadillo****:* ***Rcpp*** *integration for Armadillo templated linear algebra library.* R package version 0.3.2.4, URL http://CRAN.R-project.org/package=RcppArmadillo.

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 (Statistical Methodology)*, **66**(1), 131–143.

Loy A (2012). ***HLMdiag****: Diagnostic tools for two-level normal hierarchical linear models.* R package version 0.1.5, URL http://CRAN.R-project.org/package=HLMdiag.

R Core Team (2012). *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/.

Sanderson C (2010). "Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments."

Zewotir T (2008). "Multiple Cases Deletion Diagnostics for Linear Mixed Models." *Communications in Statistics - Theory and Methods*, **37**(7), 1071–1084.

Zewotir T, Galpin J (2005). "Influence diagnostics for linear mixed models." *Journal of Data Science*, **3**, 153–177.

Affiliation:

Adam Loy
Department of Statistics
Iowa State University
Ames, IA 50011-1210 United States of America
E-mail: aloy@iastate.edu
URL: aloy.public.iastate.edu