5-1-2019

Recursive residuals for linear mixed models

Ahmed Bani-Mustafa

K. M. Matawie

Caroline F. Finch
Edith Cowan University

Amjad Al-Nasser

Enrico Ciavolino

10.1007/s11135-018-0814-6

Recursive Residuals for Linear Mixed Models

Abstract

This paper presents and extends the concept of recursive residuals and their estimation to an important class of statistical models, Linear Mixed Models (LMM). Recurrence formulae are developed and recursive residuals are defined. Recursive computable expressions are also developed for the model's likelihood, together with its derivative and information matrix. The theoretical framework for developing recursive residuals and their estimation for LMM varies with the estimation method used, such as the fitting-of-constants or the Best Linear Unbiased Predictor (BLUP) method. These methods are illustrated through application to an LMM example drawn from a published study. Model fit is assessed through a graphical display of the developed recursive residuals and their Cumulative Sums (CUSUM).

Keywords: BLUP; Fitting-of-Constant; Linear Mixed Model; Recursive Estimation; Recursive Residuals.

1. Introduction

Recursive residuals are useful and powerful analysis tools for a wide variety of fixed effect models, particularly in providing diagnostic tests for detecting serial correlation, heteroscedasticity, functional misspecifications and structural change in regression models. Together with estimation of the model parameters they have the best statistical properties (including independency) and provide intuitive graphical tools for investigating changes of model parameters overtime using the Cumulative Sum (CUSUM) test. Therefore, recursive residuals and estimates have been used in many areas of application and have proved useful diagnostic tools in regression model checking.[1-5]

Recursive residuals and estimation for linear regression models with independent errors were first introduced by Plackett [6] and included into a set of diagnostic tests by Brown et al. [7]. Since then, the concepts have been applied to dependent error models by McGilchrist et al. [8] and to repeated measures analysis by McGilchrist and Cullis [9]. Tobing and McGilchrist [10] derived formulae for recursive estimation of unknown parameters and the vector of recursive residuals for multivariate models. McGilchrist and Matawie [11] introduced an extension of recursive residuals and estimation to Generalised Linear Models (GLM). Although, General and GLM diagnostic tools and the process of checking their components have been discussed using recursive residuals and estimates, these approaches have not yet been applied to model with both fixed and random effects.

The theory of Linear Mixed Models (LMM), which includes both fixed and random effects, is widely used when modeling correlated outcomes. This class of

statistical models is not only used directly in many application fields but is also used as the basis of iterative steps when fitting other types of mixed-effects models, such as Generalised Linear Mixed Models (GLMM).[12]

Despite the widespread popularity of LMM, diagnostic methodology for addressing model adequacy and validity are relatively underdeveloped and the consequences of misspecifying assumptions of the LMM are not well known.[13] Verbeke & Molenberghs [14] noted that the choice of the diagnostic method for LMM is not obvious, and Agresti [15] argued that there is a lack of adequate research regarding model checking and diagnostics for mixed models in general. Lin et al. [16] proposed graphical techniques for assessing the adequacy of the deterministic portion of GLMM, but their methods do not address the random component of model fit. Recently, Houseman et al. [17] used Cholesky Residuals for assessing normal errors in LMM by supplying appropriate bounds to normal QQ plots facilitated by asymptotic error independence. Several other authors have proposed goodness-of-fit tests for application in the mixed model setting, [18,19] but their approaches are complex and do not readily lend themselves well to graphical displays.

For models with subject-specific random effects, Pinheiro & Bates [20] advocated the use of standardized residuals formed when using predictions of subject-specific means and an estimate of residual error. However, comparison of only fitted and observed values can be misleading, as such comparisons reflect intended shrinkage of estimates towards the overall mean. Hilden-Minton [21] proposed the least confounded residuals, which depend only on fixed components and on the error that it is supposed to predict. Nobre & Singer [22] presented formulae for calculating studentized subject-specific residuals for linear mixed and suggested to use of the normal quantile graph plots with simulated envelops to check the assumption of normality.

Given the above limitations and gaps in the published literature, there is a need to develop recursive residuals that facilitate independency for LMM as a new important and powerful diagnostic tool to the LMM. Such tools (based on the recursive residuals) would have the best properties in checking the model components and validity, since the recursive residuals behave exactly as under the null hypothesis, until change in the model occurs.[7]

The approaches we consider in this paper for the development of recursive residuals and their estimates for LMM are based on well-known LMM estimation methods, such as Henderson's fitting-of-constants method and the Best Linear Unbiased Predictor (BLUP) method. Henderson's [24] fitting-of-constants method, which is also known as the Ordinary Least Squares (OLS) method, is used extensively in Analysis of Variance (ANOVA). This approach replaces the Sum of Squares in a balanced ANOVA by quadratic forms involving Least Squares solutions of effects from which parameters and variances are to be estimated. In section 2, we will define and derive computation formulae for the recursive residuals and estimates using the fitting-of-constants method.

Because the derivation of recursive residuals and estimates for LMM using the BLUP estimation method is not as tranquil as the derivation of recursive residuals for OLS, the formulation of Lee & Nelder [25] for the fixed and the random design matrices of the LMM are used instead. In section 3, we will introduce and discuss the LMM recursive estimations and residuals formulation.

Most of the relevant literature presents two methods for calculating recursive residuals and recursive estimates. The first method is based on setting the initial estimates and the corresponding matrices to null (Natural Order). The data are then entered progressively to estimate and update the already estimated parameters and to calculate

the recursive residuals when they become available. This method of calculation has been used by various researchers including McGilchrist and colleagues, in developing recursive procedures for different models.[8,11,25,26]

The second method of calculating recursive residuals is based on fitting an initial number of observations, considered as a base, which are needed to estimate all the model parameters. When the base is fitted, the remaining observations are entered progressively. The estimates are then updated and recursive residuals are produced. Entering the remaining observations progressively can be done either in ascending order, which is called forward recursive residuals, or in descending order, leading to backward recursive residuals. [1,27,28]

In this paper, both methods will be used for calculating the recursive residuals and updating the parameter estimates for LMM. Our development and presentation of the recursive residuals and estimates will be therefore largely based on McGilchrist et al. [8] notations and formulations. Recursive computable expressions for the likelihood function and its derivative and information matrix will be obtained and given in sections 2 and 3. Finally, recursive residuals and estimation for both methods, OLS and BLUP, are applied to the same example in section 4.

2. OLS Recursive Residuals for LMM's

Recursive estimation is a technique for updating parameter estimates where the resulting change in the estimates is proportional to the recursive residuals. The recursive residual corresponding to an observation Y_t at time t, is the scaled difference between Y_t and its best predictor using observations recorded prior to time t. Thus, current and successive predictors of Y_t are computed recursively based on parameter estimates from observations prior to t. In this section, the approach used for developing recursive residuals and estimates of LMM is based on Henderson's [24] fitting-of-constants method and McGilchrist et al. [8] notations and formulations.

Let Y_t be the continuous observation on the dependent variable at time t corresponds to vectors of regression variables $\boldsymbol{x}_t$ and $\boldsymbol{s}_t$. The LMM we consider at time t can be expressed as

$$Y_t = \boldsymbol{x}_t\boldsymbol{\beta} + \boldsymbol{s}_t\boldsymbol{u} + E_t, \qquad t = 1, 2, \dots, n \tag{1}$$

where at time t there are t observations (the first t observations) in the following matrices and vectors and the remaining observations $n - t$ are considered zeros (i.e., $t + 1, t + 2, \dots, n$)

$\boldsymbol{y}'_{\boldsymbol{n}} = [Y_1, Y_2, \dots, Y_n]$: $\boldsymbol{y}$ is the response vector $n \times 1$,

$\boldsymbol{X}'_{\boldsymbol{n}} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]$: $\boldsymbol{X}$ is the observed design matrix for the fixed effect $n \times p$ matrix,

$\boldsymbol{\beta}' = [\beta_1, \beta_2, \dots, \beta_p]$: $\boldsymbol{\beta}$ is the unobserved parameter vector of fixed effects $p \times 1$,

$\boldsymbol{S}'_{\boldsymbol{n}} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n]$: $\boldsymbol{S}$ is the observed design matrix for the random effect $n \times r$,

$\boldsymbol{u}' = [u_1, u_2, \dots, u_r]$: $\boldsymbol{u}$ is the vector of unobserved random effect $r \times 1$, with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Var(\boldsymbol{u}) = \sigma_{\boldsymbol{u}}^2\boldsymbol{I} = \boldsymbol{G}$

$\boldsymbol{\varepsilon'_t} = [E_1, E_2, \dots, E_n\,]$: $\boldsymbol{\varepsilon}$ is the error term vector $n \times 1$, assumed to be independent and normally distributed with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \boldsymbol{I} = \boldsymbol{R}$.

Assume that all levels of $\boldsymbol{u}$ pertain to the same source of variation, such that $Var(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I} = \boldsymbol{G}$ and $Cov(\boldsymbol{u}; \boldsymbol{\varepsilon}) = \mathbf{0}$.

The recursive estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}})$ is the solution of Henderson's equations (2) or the OLS equations for the model in (1) which is based on treating $\boldsymbol{u}$ and $\boldsymbol{\beta}$ as fixed effects at any time t.

$$\begin{bmatrix} \boldsymbol{X}_t'\boldsymbol{X}_t & \boldsymbol{X}_t'\boldsymbol{S}_t \\ \boldsymbol{S}_t'\boldsymbol{X}_t & \boldsymbol{S}_t'\boldsymbol{S}_t \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}_t \\ \widehat{\boldsymbol{u}}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t'\boldsymbol{y}_t \\ \boldsymbol{S}_t'\boldsymbol{y}_t \end{bmatrix} \tag{2}$$

Where the $\boldsymbol{X}_t, \boldsymbol{S}_t, \boldsymbol{y}_t$ are the design matrices at time t.

The recursive residual W_t and recursive estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ can be written as (see the full derivation in [29])

$$W_t = \begin{cases} c_t^{*-1/2}\left[Y_t - \boldsymbol{z}_t'\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{pmatrix}\right] & \text{rank stays same} \\ 0 & \text{rank increases.} \end{cases} \tag{3}$$

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_t \\ \widehat{\boldsymbol{u}}_t \end{bmatrix} = \begin{cases} \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{bmatrix} + c_t^{*-1/2} W_t \boldsymbol{g}_t^* & \text{rank stays same} \\ \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{bmatrix} + c_t^{\dagger-1}\left[Y_t - z_t'\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{pmatrix}\right]\boldsymbol{g}_t^{\dagger}, & \text{rank increases.} \end{cases} \tag{4}$$

As data becomes more available and observations are added recursively, the design matrices $\boldsymbol{X}_{t-1}, \boldsymbol{S}_{t-1}, \boldsymbol{y}_{t-1}$ are replaced by $\boldsymbol{X}_t, \boldsymbol{S}_t, \boldsymbol{y}_t$ and the rank of $\boldsymbol{H}_t$ may increase by one or remains the same. A test for the rank change is to consider using the following vector

$$\boldsymbol{q}_t = \boldsymbol{z}_t'(\boldsymbol{I} - \boldsymbol{H}_{t-1}^{-1}\boldsymbol{H}_{t-1}\,).$$

The test vector may result with a zero vector (rank stays the same) or contain at least one non-zero element which is used to choose the corresponding component of the vector of independent variables $\boldsymbol{x}_t$ and $\boldsymbol{s}_t$ and accordingly estimate its recursive regression coefficient.

So if the rank stays the same then the following equations are used in calculating the recursive estimates in equation (4) and the recursive residual will be 0.

$$c_t^* = 1 + \boldsymbol{z}_t'\,\boldsymbol{H}_{t-1}^{-1}\boldsymbol{z}_t \qquad \boldsymbol{z_t} = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{s}_t \end{bmatrix}$$

$$\boldsymbol{g}_t^* = \boldsymbol{H}_{t-1}^{-1}\boldsymbol{z}_t \qquad \boldsymbol{H}_t^{-1} = \boldsymbol{H}_{t-1}^{-1} - c_t^{*-1}\boldsymbol{g}_t^*\boldsymbol{g}_t^{*\prime}$$

In the same way, the following equations are used to calculate the recursive residuals in (3) and the recursive estimates in (4) if the rank increases by one

$$c_t^{\dagger} = \boldsymbol{z}_t \boldsymbol{g}_t^{\dagger} \qquad \boldsymbol{g}_t^{\dagger} = (\boldsymbol{I} - \boldsymbol{H}_{t-1}^{-1}\boldsymbol{H}_{t-1})$$

$$\boldsymbol{v}_t = \boldsymbol{g}_t^{\dagger} - c_t^{\dagger} c_t^{*-1} \boldsymbol{g}_t^{*} \qquad \boldsymbol{H}_t^{-1} = \boldsymbol{H}_{t-1}^{-1} + c_t^{*-1} \boldsymbol{g}_t^{*} \boldsymbol{g}_t^{*\prime} + c_t^{\dagger -2} c_t^{*} \boldsymbol{v}_t \boldsymbol{v}_t'$$

There are two methods to calculate the recursive residuals and estimates. First, the Natural Order (NO) method and fitting initial base. NO method starts at time $t = 0$ with a zero vector as an initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$, for, $\boldsymbol{H}_t = \boldsymbol{0}$, its inverse $\boldsymbol{H}_t^{-1}$ and their product $\boldsymbol{H}_t^{-1}\boldsymbol{H}_t$ are taken to $(p + r) \times (p + r)$ matrices of zeros. As observations are added recursively, the design matrices $(\boldsymbol{X}_{t-1}, \boldsymbol{S}_{t-1}, \boldsymbol{y}_{t-1})$ are updated with the new observation $(\boldsymbol{X}_t, \boldsymbol{S}_t, \boldsymbol{y}_t)$ and the rank of $\boldsymbol{H}_t$ may remain the same or increase by one.

The second method of calculating the recursive residuals starts with estimating the model parameters using initial base of $(p + r - 1)$ observations. This initial base should be selected in a way to provide a non-singular information matrix in which all parameters are estimable. After estimating the model parameters observations are entered progressively and update (fine tune) the parameters and estimate $(n - p - r + 1)$ recursive residuals. [1,27]

3. BLUP Recursive Residuals

In this section the recursive residuals and estimates for LMM are developed based on the BLUP method. The BLUP estimation method is based on maximising the sum of the log-likelihood function, which is a penalised likelihood function. The likelihood estimators of $\boldsymbol{\beta}$ is the solution of the Mixed Model Equations (MME); which is given in the following form [31]:

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{S} \\ \boldsymbol{S}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{S}'\boldsymbol{R}^{-1}\boldsymbol{S} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{S}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

where $\boldsymbol{R} = \sigma_{\boldsymbol{\varepsilon}}^2 \boldsymbol{I}$ and $\boldsymbol{G} = \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}$. This solution can be simplified to:

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{S} \\ \boldsymbol{S}'\boldsymbol{X} & \boldsymbol{S}'\boldsymbol{S} + \lambda\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{S}'\boldsymbol{y} \end{bmatrix},$$

where $\lambda = \sigma_{\boldsymbol{\varepsilon}}^2 / \sigma_{\boldsymbol{u}}^2$ is the precision ratio.

The MME equations are slightly different to the OLS equations. MME includes the term $\lambda\boldsymbol{I}$ resulting from the random component. The resulting MME cannot be solved in the same way used for OLS equations. So Lee and Nelder [25] formulation for the fixed and random design matrices, $\boldsymbol{X}$ and $\boldsymbol{S}$ are used to formulated the MME. Formulated MME can be written as

$$\begin{bmatrix} \boldsymbol{X}'^{*}\boldsymbol{X}^{*} & \boldsymbol{X}'^{*}\boldsymbol{S}^{*} \\ \boldsymbol{S}'^{*}\boldsymbol{X}^{*} & \boldsymbol{S}'^{*}\boldsymbol{S}^{*} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'^{*}\boldsymbol{y} \\ \boldsymbol{S}'^{*}\boldsymbol{y} \end{bmatrix}, \tag{5}$$

where

$$\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{0} \end{pmatrix}, \qquad \boldsymbol{S}^* = \begin{pmatrix} \boldsymbol{S} \\ \sqrt{\lambda}\boldsymbol{I} \end{pmatrix}, \qquad \boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ E(\boldsymbol{u}) \end{pmatrix},$$

where $\boldsymbol{0}$ is the $r \times p$ zero matrix and $\boldsymbol{I}$ is the $r \times r$ identity matrix. Formulating the design matrices is also called a pseudo data approach [20].

It can be shown that the BLUP recursive residuals and estimates for MME equations in (5) are

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_t \\ \widehat{\boldsymbol{u}}_t \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{bmatrix} + c_t^{*-1/2} W_t \boldsymbol{g}_t^*, \qquad t = p + r, \dots, n \tag{6}$$

and the individual recursive residual is

$$W_t = c_t^{*-1/2} \left[Y_t - \boldsymbol{z}_t' \begin{pmatrix} \widehat{\boldsymbol{\beta}}_{t-1} \\ \widehat{\boldsymbol{u}}_{t-1} \end{pmatrix} \right]. \tag{7}$$

For more details for the recursive estimation derivation see [29].

In the same way illustrated in Section 2, both methods of recursive estimation can be used Natural Order (NO) and initial base using the new design matrices $\boldsymbol{X}_t^*, \boldsymbol{S}_t^*, \boldsymbol{y}_t^*$.

4. Example

To illustrate the application and computation of these developed formula for recursive residuals and estimates for LMM we used Nobre & Singer [22] data. The data relates to a comparison of the capacity to remove bacterial plaque with continuous daily use with a low cost monoblock toothbrush against a conventional toothbrush. Indices of plaque in 32 children (aged 4 – 6 years) were measured before and after tooth brushing at four evaluation sessions. The data is an example of repeated measurements, taken on the same experimental units over four evaluation sessions, adjusting for pretreatment bacterial plaque indices.

Nobre & Singer [22] fitted the following LMM:

$$\ln y_{ijd} = \alpha_j + \beta \ln x_{ijd} + b_i + \varepsilon_{ijd},$$

where $y_{ijd}(x_{ijd})$ is the post-treatment (pre-treatment) bacterial plaque index for the i^{th} subject evaluated in the d^{th} session with the j^{th} type of toothbrush, α_j is the fixed effects associated with the two types of toothbrush, β is a pretreatment bacterial plaque index coefficient. The $b_i \sim N(0, \tau^2)$ are the subject random effects, $\varepsilon_{ijd} \sim N(0, \sigma^2)$ are the random measurement errors and $cov(\boldsymbol{b}, \boldsymbol{\varepsilon}) = \boldsymbol{0}$.

Our reframing of this LMM in a matrix form at time t as:

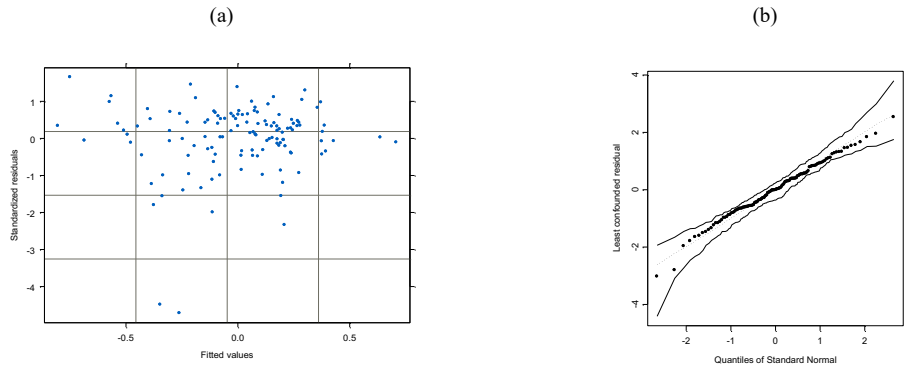
$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{S}_t \boldsymbol{u} + \boldsymbol{\varepsilon}_t, \qquad t = 1, 2, \dots, 128 \tag{9}$$

where matrices and vectors at time t has only the first t observations and the remaining

observations $(n-t)$ are zeros with the same dimensions defined in (1). $\boldsymbol{y}$ is a 128×1 response vector of post-treatment bacterial plaque indices (y_{ijd}), $\boldsymbol{X}$ is a 128×3 fixed effects design matrix (intercept, two types of toothbrush and log(pre-treatment)) and $\boldsymbol{S}$ is a 128×32 random effects design matrix (subject effect), $\boldsymbol{\varepsilon}$ is a 128×1 errors vector normally distributed with zero mean and $\sigma_{\varepsilon}^2 \boldsymbol{I}$ variance. The vector $\boldsymbol{\beta}$ is a 3×1 vector of fixed effects that are unknown and $\boldsymbol{u}$ is a 32×1 vector of random effects normally distributed with zero mean and $\sigma_{\boldsymbol{u}}^2 \boldsymbol{I}$ variance and $\boldsymbol{u} \perp \boldsymbol{\varepsilon}$.

The LMM obtained (using the `S-plus lme` package) was significant, with *t*-values of $(-10.027,\ 16.087,\ 2.98)$ for the intercept, pretreatment and treatment effects respectively. The LMM appeared to be satisfactory. A plot of the standardized residuals $(\hat{\boldsymbol{\varepsilon}}/\hat{\sigma}_e)$ versus the fitted values, shown in Figure 1(a), was satisfactory except for an indication of a possible two outliers. Hilden-Minton [21] proposed the least confounded residuals, which depends only on fixed components and on the error that it is supposed to predict. Nobre & Singer [22] presented formulae for calculating studentized subject-specific residuals for linear mixed and suggested to use of the normal quantile graph plots with simulated envelops to check the assumption of normality. The approach has been developed further developed further by Schützenmeister and Piepho (2010) [23]. Both graphs are given in Figure 1. The normal quantile plot for the residuals (Figure 1(b)) did not identify any observation outside the simulated envelope without any trend, suggesting normality of the residuals. These results are similar to those given by Nobre & Singer [22] who investigated three techniques of residuals analysis for LMM using the same data.

Figure 1. Standardized residuals (a) and simulated 95% confidence envelope for the standardized least confounded residuals (b) for model (5.1)



Calculating the parameters of the above model involves the solution of a set of $p+r$ simultaneous linear equations. Often in practice, we do not have access to all the data required to give an estimate of Y. One advantage of the recursive estimation is to provide a way of getting the estimate of Y which is constantly updated by data arriving. Parameters of the above model can be calculated using the first $(p+r-1)$ observations that corresponds to each fixed and random effects, then fine-tuned as more observations become available. Recursive estimation allows for tracking of the value of the parameters over time and to check for structural breaks. It should be apparent that the overall result of these successive iterations agrees with the one obtained if we process all data after they have been collected. The advantage of the recursive estimation is that at each stage we have the best representation of what we know about the parameters despite the size of data.

The recursive residuals and their corresponding CUSUM are useful and powerful tools for detecting and checking the functional misspecifications, model validity and stability. In a well–specified model, the recursive residuals would have a mean of zero ($\sum_{t=1}^{n} W_t = 0$). If all regression assumptions are satisfied, the CUSUM plot should show a random walk within a parabolic envelope about the origin, since the expectation of these recursive residuals is zero.[1]

Furthermore, it is preferable to use recursive residuals rather than ordinary residuals to detect a change in the model, since the recursive residuals behave exactly as under the null hypothesis (the cumulative sum of RR will have a sum of zero), until a change in the model occurs.[7]

In the case where a model misspecification arises, under common circumstances the recursive residuals will have a mean non-zero difference and it is possible to detect changes over times that are sustained beyond some specific point of time. Recursive residuals may also be used to depict the pattern of variation of the proposed model. Note that recursive residuals are not used to model variation over time, but rather to depict the type of variation from a stationary model that is generated from the data.

The purpose of this paper is not to discuss the use of recursive residuals for model improvement; rather it shows how they may be defined and provides an illustration of their use. The major benefit of using recursive residuals and their corresponding CUSUM, is the power in detecting change over time that is sustained beyond some specific time point. The precise method of using recursive residuals depends on what change is suspected to have occurred and hence varies with the application. Such model improvements may be suggested by the recursive residuals. For more details on model improvement using recursive residuals, see [1,27,32,33].

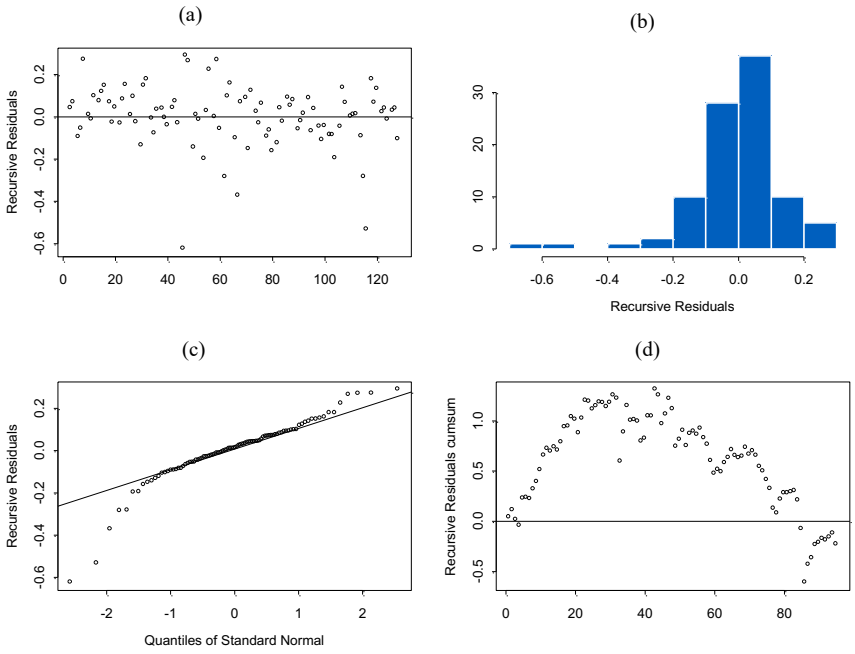
4.1. OLS Recursive Residuals

The OLS recursive estimation procedures developed in the previous sections were programmed in `S-plus`. The recursive residuals for the LMM in (9) were calculated starting with zero as initial estimates of $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{u}}_0$ and related matrices such as $\boldsymbol{H}_0$ and $\boldsymbol{H}_0^{-1}$. This method of calculation is adopted as it is more general and practical than fitting an Initial Base. With the addition of each new observation, updated estimates of $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are obtained by using the recursive algorithm summary given in the previous section. As mentioned above, recursive residuals occur whenever the rank of the information matrix does not increase. Therefore, the number of recursive residuals calculated here will be $(n - rank(\boldsymbol{X}) + 1 - rank(\boldsymbol{S}) + 1 = 95)$ after estimating two fixed effects ($rank(\boldsymbol{X})$) and predicting 31 mixed effects factors ($rank(\boldsymbol{S}) - 1$) random effects, since $E(\boldsymbol{u}) = 0$)). The recursive residuals and their CUSUM and normal quartile plots appear in Figures 2(a-d). The CUSUM at observation i is $C_i = \sum_{t=1}^{i} W_t$ where W_t is the recursive residual at observation t.

The recursive residuals Figure (2a) is very close to the standardized residuals plot (Figure 1(a)), and is considered satisfactory with a potential of two outliers. The normal probability plot (Figure (2c)) shows an approximately straight line with the data points at the two ends of the line indicating possible low values and/or outliers. However, the CUSUM plot (Figure (2d)) shows an initial upward trend followed by a downward trend indicating some sort of model misfit with a negative CUSUM ($\sum W_t = -0.231$). Model misfit such as this may be due to various reasons including outliers, an omitted variable, incorrect model specification or incorrect model underlying distributional assumptions. For more details see [1,27,32]. Dealing with model misfit may improve the above model, but this is not the aim of this example.

It should pointed out that in their initial data paper, Nobre & Singer [22], also identified the same two outliers, which were identified by standardized residuals and recursive residuals plots shown above. Removing these two outliers did not improve the fit of the model or CUSUM plot.

Figure 2. OLS recursive residuals (a), Histogram (b), Normal Quantile-Quantile plot (c) and CUSUM graph (d) for fitting-of-constants Recursive Residuals

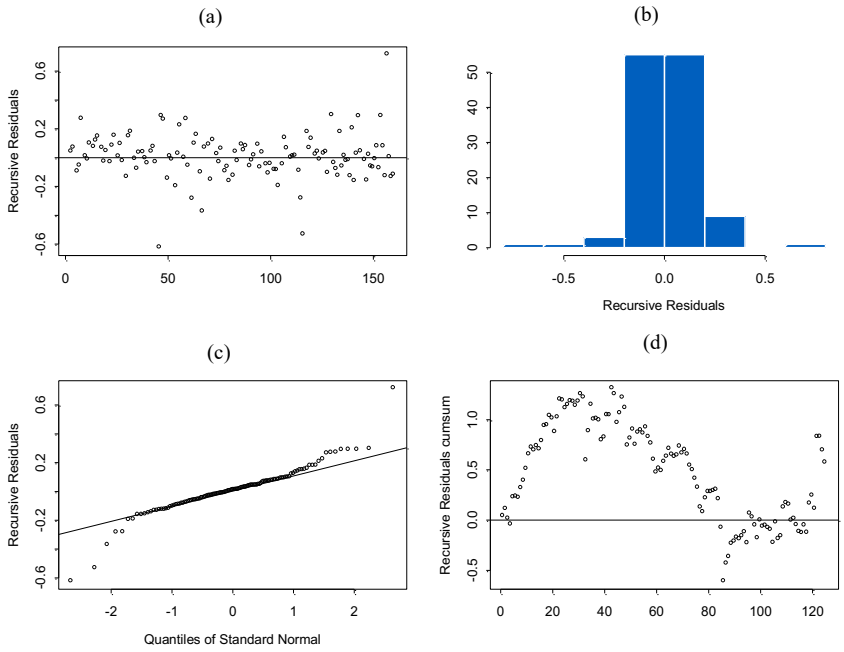


4.2. BLUP Recursive Residuals

The BLUP recursive estimates and residuals were obtained for the same model (5.1) with adjustments for this data case. The initial estimates of $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{u}}_0$ and related matrices such as $\boldsymbol{H}_0$ and $\boldsymbol{H}_0^{-1}$ were set to null. The data were then entered progressively to estimate and update the already estimated parameters and the recursive residuals calculated when they became available. The number of recursive residuals calculated here will be $(n - rank(\boldsymbol{X}) = 125)$ after estimating three fixed effects $(rank(\boldsymbol{X}))$. The recursive residuals and their CUSUM and normal quartile plots appear in Figures 3(a-d).

The recursive residuals Figure 3(a) is satisfactory. However, it is slightly different from the OLS recursive residuals and standardized residuals in that it shows the potential for three outliers. The normal probability plot (Figure 3(c)) shows an approximately straight line with a potential for outliers with the data points at the two ends of the line indicate low values and/or outliers. Importantly, the CUSUM plot shows an initial upward trend followed by a downward trend followed by a random segment before an upward trend again. This is an indicator of model misfit with a positive CUSUM ($\sum W_t = 0.572$). The BLUP recursive residuals show less variability than the OLS recursive residuals around the zero line and the CUSUM plot also becomes more stable at the end. The reason for this might be that the contribution of the random effect in the model using BLUP recursive method is more than the contribution of the random effect in the OLS recursive method, i.e., in the BLUP method the random effects explains more variation in the model than what can be achieved by OLS recursive method.

Figure 3. BLUP Recursive Residuals (a), Histogram (b), Normal Quantile-Quantile plot (c) and CUSUM graph (d) for BLUP Recursive Residuals



4.3. OLS and BLUP Recursive Residuals

The main purpose of using this particular example is to demonstrate and elaborate on how the recursive procedures for LMM are developed and applied. Thereafter, these recursive residuals can be incorporated into a set of diagnostic tests such as the CUSUM, signed residuals and their CUSUM, normal probability plots, cumulative sums of the square roots of the absolute value of the standardized recursive residuals and cumulative sums of squares (CUSUMSQ) of recursive residuals.[11]

In this example, both OLS and BLUP recursive residuals analyses showed a slight model misfit and also identified the outliers found by Nobre & Singer [22] in their initial analysis of data. Our example shows that, in this case, the OLS residuals undervalue the data ($\sum W_t = -0.231$) and BLUP recursive residuals overvalue the data ($\sum W_t = 0.572$). One might try to remedy this problem by stepwise selection from higher order polynomial terms, including more relevant variables in the model or replacing the normality assumption for the random effects distribution by a more robust heavy tails distribution such as *t*-distribution.

Despite the model misfit, the BLUP recursive residuals showed less variability than the OLS recursive residuals. This highlights the importance of the random effects for the model in this instance in explaining the variation in the data. Random effects were found to be significant using Akaike information criterion (AIC) in `S-plus`. This also explains how the OLS recursive residuals can be used to demonstrate the importance of including the random effects to a fixed effect model. The BLUP recursive residuals showed a potential third outlier which was not detected by the OLS recursive residuals or by Nobre & Singer [22].

5. Discussion

In summary, this article has newly defined and derived recursive residuals and estimates using two estimation techniques for LMM: the OLS and BLUP estimation methods. Previously, LMM diagnostic tools have been relatively underdeveloped and as a result, the consequence of misspecifying assumptions of LMM not well known. We have presented and illustrated that recursive residuals can be usefully incorporated into a set of diagnostic tests as they behave exactly as under the null hypothesis, until a change in the model occurs. The resultant diagnostic and graphical tests (based on recursive residuals), along with their behavior and properties, can be investigated for the LMM under different functional misspecification.

The BLUP recursive residuals treat random effects as random, while the OLS method treats them as fixed. The reason for including the OLS recursive residuals and estimation are the extensive use of this estimation technique in ANOVA. The OLS recursive residuals and estimates could be incorporated into a set of diagnostic tests to detect the importance of the random effects (LMM) in modeling the data. The BLUP showed more stable recursive residuals than those generated by OLS, supporting the significance of the random effects in the model.

In conclusion, recursive residuals perform well in model fit diagnostics and provide more information about model misspecification than using ordinary residuals. It should now be possible to extend the development of recursive residuals and estimates technique for a more general and Generalised Linear Mixed Models.

All calculations are implemented in the `S-plus (lme)` and `R (nlme and lmer)` packages with modification. The codes employed for calculating Recursive Residuals are developed in `S-plus` and `R` and can be obtained directly from the authors.

Acknowledgements: This research was supported in part by an Early Career Researcher Grant (ECR); National Health and Medical Research Council (NHRMC) Principle Research Fellowship, Federation University Australia, Ballarat, Australia.

References

[1] Galpin J, Hawkins D. The use of recursive residuals in checking model fit in linear regression. American Statistician. 1984; 38:94-105.

[2] Hawkins D, Olwell D. Cumulative sum charts and charting for quality improvement. New York: Springer Verlag;1998.

[3] Does R, Koning A. CUSUM charts for preliminary analysis of individual observations. Journal of Quality Technology. 2000;32:122–132.

[4] Haslett J, Haslett S. The three basic types of residuals for a linear model. International Statistical Review. 2007; 75(1):1-24.

[5] Godolphin J. New formulations for recursive residuals as a diagnostic tool in the fixed-effects linear model with design matrices of arbitrary rank. Computational Statistics and Data Analysis. 2009;53(6):2119-2128.

[6] Plackett R. Some theorems in least squares. Biometrika. 1950; 37(1-2):149-157.

[7] Brown R, Durbin J, Evans J. Techniques for testing the constancy of regression relationships over time Journal of the Royal Statistical Society Series B (Methodological). 1975;149-192.

[8] McGilchrist C, Sandland R, Hennessy J. Generalized inverses used in recursive residuals estimation of the general linear model. Australian & New Zealand Journal of Statistics. 1983;25(2):321-328.

[9] McGilchrist C, Cullis B. REML estimation for repeated measures analysis. Journal of Statistical Computation and Simulation. 1991;38(1):151-163.

[10] Tobing H, McGilchrist C. Recursive residuals for multivariate regression models. Australian & New Zealand Journal of Statistics. 1992; 34(2):217-232.

[11] McGilchrist C, Matawie K. Recursive residuals in generalised linear models. Journal of Statistical Planning and Inference. 1998;70(2):335-344.

[12] Bates D, DebRoy S. Linear mixed models and penalized least squares. Journal of Multivariate Analysis. 2004;91(1), 1-17.

[13] Jacqmin-Gadda H, Sibillot S, Proust C, Molina J, Thiebaut R. Robustness of the linear mixed model to misspecified error distribution. Computational Statistics and Data Analysis. 2007;51(10):5142-5154.

[14] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer-Verlag; 2000.

[15] Agresti A. Categorical data analysis. 2nd ed. Wiley: New York; 2002.

[16] Lin D, Wei L, Ying Z. Model-checking techniques based on cumulative residuals. Biometrics. 2002;58:1-12.

[17] Houseman E, Ryan L, Coull B. Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. Journal of the American Statistical Association. 2004;99(466):383-394.

[18] Hodges J. Some algebra and geometry for hierarchical models, applied to diagnostics. Journal of the Royal Statistical Society Series B, Statistical Methodology. (1998);60: 497-536.

[19] Jiang J. Goodness-of-fit tests for mixed model diagnostics. Annals of Statistics. 2001;29(4):1137-1164.

[20] Pinheiro J, Bates D. Mixed-effects models in S and S-PLUS. NY Springer; 2000.

[21] Hilden-Minton JA. Multilevel diagnostics for mixed and hierarchical linear models [dissertation]. University of California, Los Angeles; (2001).

[22] Nobre J, Singer J. Residual analysis for linear mixed models. Biometrical journal. Biometrische Zeitschrift. 2007;49(6):863-875.

[23] Schützenmeister, A., Piepho, H., 2012. Residual analysis of linear mixed models using a simulation approach. Computational Statistics and Data Analysis 56, 1405- 1416.

[24] Henderson C. Estimation of variance and covariance components. Biometrics. 1953;226-252.

[25] Lee Y, Nelder J. Hierarchical generalized linear models. Journal of the Royal Statistical Society Series B (Methodological). 1996;58(4):619-678.

[26] McGilchrist C, Sandland R. Recursive estimation of the general linear model with dependent errors. Journal of the Royal Statistical Society Series B (Methodological). 1979;65-68.

[27] McGilchrist C, Lianto S, Byron D. Signed residuals. Biometrics. 1989;237-246.

[28] Kianifard F, Swallow W. A review of the development and application of recursive residuals in linear models. Journal of the American Statistical Association. 1996; 91:391-400.

[29] Ahmed Bani-Mustafa, Recursive Residuals for Linear Mixed Models. PhD thesis. University of Western Sydney, Australia; 2004.

[30] Pickford M, Haslett S. A statistical test of single firm market power. New Zealand Economic Papers. 1999;33:39–58

[31] Bartlett M. An inverse matrix adjustment arising in discriminant analysis. The Annals of Mathematical Statistics. 1951;107-111.

[32] Henderson CR. Estimation of variances and covariances under multiple trait models. J. Dairy Sci. 1984;67:1581-1589

[33] Hawkins D. Diagnostics for use with regression recursive residuals. Technometrics. 1991;33(2):221-234.

[34] de Luna X, Johansson P. Testing exogeneity under distributional misspecification. Working Paper: 2001;9. Institute for Labour Market Policy Evaluation (IFAU), Uppsala.

[35[Khamis, F. (2017). Crime and divorce. Can one lead to the other? Using Multilevel Mixed Models. *Electronic Journal Of Applied Statistical Analysis, 10*(2), 328 - 348.