

Update of the nlme Package to Allow a Fixed Standard Deviation of the Residual Error

by Simon H. Heisterkamp, Engelbertus van Willigen, Paul-Matthias Diderichsen, John Maringwa

Abstract The use of linear and non-linear mixed models in the life sciences and pharmacometrics is common practice. Estimation of the parameters of models not involving a system of differential equations is often done by the R or S-Plus software with the nonlinear mixed effects **nlme** package. The estimated residual error may be used for diagnosis of the fitted model, but not whether the model correctly describes the relation between response and included variables including the true covariance structure. The latter is only true if the residual error is known in advance. Therefore, it may be necessary or more appropriate to fix the residual error a priori instead of estimate its value. This can be the case if one wants to include evidence from past studies or a theoretical derivation; e.g., when using a binomial model. S-Plus has an option to fix this residual error to a constant, in contrast to R. For convenience, the **nlme** package was customized to offer this option as well. In this paper, we derived the log-likelihoods for the mixed models using a fixed residual error. By using some well-known examples from mixed models, we demonstrated the equivalence of R and S-Plus with respect to the estimates. The updated package has been accepted by the Comprehensive R Archive Network (CRAN) team and will be available at the CRAN website.

Introduction

Life science and pharmacometric studies almost always involve an application of linear (Laird and Ware, 1982) and non-linear (Davidian and Giltinan, 1995) mixed models. In pharmacometrics, the use of nonlinear mixed effects modelling (NONMEM) software (Beal and Sheiner, 1988) for models with systems of differential equations is predominant. Simpler pharmacokinetic (PK) or pharmacodynamic (PD) models may be based on curve fitting, for which software such as S-Plus and R are preferred. Both programs use a similar mixed model package, respectively, **nlme** and nlme library, both of which were originally developed by the authors Pinheiro and Bates (Pinheiro and Bates, 2001). In most cases, the estimates of the parameters are the same at least to the third significant digit, although the same model and data might occasionally lead to convergence problems (e.g., in R and not in S-Plus, and vice versa), although larger differences may occur e.g., when performing an analysis of variance with the function `anova`, as shown in 2.6. Typically, the estimated fixed and random coefficients have a practical interpretation. The fixed coefficients usually refer to the parameters of interest for the study; i.e. treatment, gender differences, half-life of the drug substance etc., while the random parameters are interpreted as variation in the population represented by the study. The residual error of the fitted model is an important indicator for the goodness of fit. Whether the goodness of fit is assessed by the log-likelihood ratio, the Akaike information criterion (AIC) (Akaike, 1980) or the Bayesian information criterion (BIC) (Schwartz, 1978) criteria, all criteria implicitly or explicitly use the residual error of the fit.

For some applications, it is not appropriate to estimate the residual error, especially when it is known in advance based on evidence from past studies, or a theoretically derived scaling parameter. An example of the former are meta-analyses where the standard deviation of individual patient outcomes in each study is reported. An example of the latter is a data transformation, implying a fixed scaling parameter; e.g., the square root arcsine or logit transformation may be used as an approximation of the binomial distribution.

As pointed out by an anonymous reviewer, other applications may be considered as well (Littel et al., 2006). For example simulation of new observations from the fitted model while keeping sigma constant or performing power calculations where commonly the expected error is assumed to be known and kept constant. In both cases one must bear in mind that the simulated data will not cover the whole range of possible outcomes, and in case of power calculation the actual power may be well below the nominal power one aimed at.

The fitting functions from the **nlme** library in S-Plus allow the residual standard error to be fixed and kept constant during the fitting process. This feature was not described in Pinheiro and Bates (Pinheiro and Bates, 2001) but has been added as an argument to the respective control functions. This feature was not available in the R package **nlme** but is included from version 3.1.123 (released 2016-01-17) onward.

We present this feature of the control functions used in linear mixed effects `lme`, non-linear mixed

effects `nlme`, generalized linear least squares `gls` and generalized non-linear least squares `gnls`. Utilities as `summary`, `anova`, `print` and `intervals` have been adapted as well.

From a coding point of view these changes were significant, as they required, among others, a change of the likelihood functions in both the R and C source codes. This paper discusses the statistical background in Section 2. In Section 3 the similarities and differences of the output of S-Plus version 8.2 are given for selected examples from Pinheiro and Bates (Pinheiro and Bates, 2001) as well as a case study describing a model-based meta-analysis of outcomes in rheumatoid arthritis. In the last Section proposals for change in `nlme` are put forward to address some of these similarities and differences.

The general linear and linear mixed model

The mixed model is an extension of the general linear model. The extension to a mixed model allows other parameters of the general linear model to have a statistical normal distribution (Laird and Ware, 1982). These models are also known as multilevel models in different fields of application, as they describe a hierarchy of levels in the data. For example, in a clinical trial, the first level might be the trial centre, the second might be the patients within the trial centre, and the third might be the time points at which the observations are acquired for an individual patient.

The reason for such a break-down of the random variation is that, within a centre, the variation between subjects may be different than between centres; e.g., differences in environment, skills of the operators, measurement devices, or the mixture of ethnicity of the patients. Variation within patients is likely to be different and correlation of longitudinal observations within a patient may be present.

The expected outcome may be linear in the design parameters, including linear combinations of polynomials and general additive models (Wood, 2006). Similarly parameters of the non-linear models are allowed to be linear combinations (Davidian and Giltinan, 1995). General linear and mixed linear models are fitted with `gls` and `lme` respectively. The corresponding non-linear models (e.g. a Michaelis-Menten model curve) are fitted with `gnls` and `nlme`.

Mixed models have two sets of parameters. The fixed parameters are often of primary interest of the study. For example, differences in treatment or gender, the half-life of a drug or factors for which one wants to control like co-medication. The random parameters are used to describe variation within respective levels of the study and are generally not controllable. For example, variation between and within a population of patients. The random parameters might be dependent on some or all of the fixed parameters or perhaps a function of the fitted values. Common statistical analysis seeks to estimate all fixed and random parameters simultaneously using the same data without prior knowledge.

In the next subsections, we describe the mathematical model extending from the general linear to the linear mixed and non-linear mixed model with knowledge of the residual error. The notation and the terminology with regard to the number of hierarchical levels follows closely to that in Pinheiro and Bates (Pinheiro and Bates, 2001). Throughout the paper we will use Greek characters to indicate the true but unknown values of the model parameters, while the corresponding character superscripted by $\hat{\cdot}$ denotes its estimate.

General linear model

The general linear model with only the basic random level of the residuals is presented here. A comparison is made between the likelihood for the usual model with an estimated residual error and the one with a fixed value of the latter. For further details of the derivation see (Pinheiro and Bates, 2001). If there were M groups (e.g. subjects) with n_i possibly correlated observations each, the total number of observations is: $N = \sum_i^M n_i$. The model is expressed for each group by the following formula:

$$y_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad \boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 \boldsymbol{\Lambda}_i) \quad (1)$$

Here, \mathbf{X} is the $n_i \times p$ design matrix and $\boldsymbol{\beta}$ is a column vector of length p . The matrix $\boldsymbol{\Lambda}_i$, the covariance matrix of each of the subjects, is a $n_i \times n_i$ positive-definite matrix is multiplied by σ^2 . Note that in this model, we do not assume any between-groups variance; σ^2 is the same for each of the M groups. Assuming that the M groups are all statistically independent; e.g., subjects in a trial the total $N \times N$ covariance matrix $\boldsymbol{\Lambda}$ is block-diagonal and each block represents the matrix $\boldsymbol{\Lambda}_i$. If all observations within a group were statistically independent as well $\boldsymbol{\Lambda}_i$ reduced to identity matrices \mathbf{I} of the same dimensions, and the model becomes an ordinary (multiple) regression model with covariance matrix \mathbf{I} multiplied by σ^2 and can be solved by standard multiple regression software.

To solve model (1), one may reduce the model to an ordinary least squares model by transforming both sides of the equation such that the covariance matrices are again diagonal and equal to $\sigma^2 \mathbf{I}$. For

details, see Appendix .1, where it is shown that formula (1) is equivalent to the least squares model following expression:

$$y_i^* = \mathbf{X}_i^* \beta + \varepsilon_i^* \quad \varepsilon_i^* \sim N(0, \sigma^2 \mathbf{I}) \quad (2)$$

The superscript * indicates the corresponding transformed vectors and matrices. This model appears to be an ordinary least squares model, however, its solution depends on unknown parameters in Λ_i , not explicitly in the equation. Thus, conditionally on the values of Λ_i , the fixed coefficients β of the model—indicated by $\hat{\beta}$ —with the usual least squares equations are estimated, yielding:

$$\hat{\beta} = [(\mathbf{X}^*)^T \mathbf{X}^*]^{-1} (\mathbf{X}^*)^T y^*$$

The matrices Λ_i usually depend on a set of only a few parameters with unknown values, which we denote as λ . This means that the estimate $\hat{\beta}$ depends on the unknown λ as well, but to simplify the notation we write $\hat{\beta}(\lambda)$, instead of $\hat{\beta}(\hat{\lambda})$. Estimation of $\hat{\beta}(\lambda)$, implies estimation of λ as well. The method employed in the package **nlme** maximizes the profiled (minus) log-likelihood which includes λ and $\hat{\beta}$ (not β) (Pinheiro and Bates, 2001). The maximization is done for both β and λ in a stepwise manner and at each step the residual variance is estimated as:

$$\hat{\sigma}^2 = \frac{\|y^* - \mathbf{X}^* \hat{\beta}(\lambda)\|^2}{N_p}$$

The number N_p is the degrees of freedom left for this residual variance. It equals N for the maximum likelihood method of estimation and $N - p$ for the restricted maximum likelihood and the least squares method. The number p is the sum of the numbers of fixed parameters and the numbers of parameters used in the function of the covariance matrix Λ . The estimate $\hat{\sigma}^2$ is plugged into all other expressions; e.g., the standard errors of the estimates and the student tests to compare the fixed parameters and the goodness of fit criteria.

The estimation of λ , β and σ is iterative because $\hat{\lambda}$, $\hat{\beta}$ and $\hat{\sigma}^2$ are interdependent. The (minus) profiled log-likelihood of the general linear model equals :

$$lik(\lambda | y, \hat{\sigma}, \hat{\beta}) = \frac{N_p}{2} \left[\log(N_p) - \log(2\pi) - 1 - \log(\|y^* - \mathbf{X}^* \hat{\beta}(\lambda)\|^2) \right] + \frac{1}{2} \sum_{i=1}^M \log |\Lambda_i| \quad (3)$$

However, if the residual variance is fixed, the profiled likelihood will be different. To see this, one has to derive the original log likelihood function without substitution of $\hat{\sigma}^2$. This profiled likelihood now depends on σ^2 instead of $\hat{\sigma}^2$, as the former is a fixed scalar which does not enter the equations.

$$llik(\lambda | y, \sigma, \hat{\beta}) = \frac{N_p}{2} \left[-\log(2\pi) - \log(\sigma^2) \right] - \frac{\|y^* - \mathbf{X}^* \hat{\beta}(\lambda)\|^2}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^M \log |\Lambda_i| \quad (4)$$

Note that if $\hat{\sigma}^2$ is substituted in (4), the expression is exactly the same as (3). The striking difference between both profiled likelihoods is that, for known σ , the function (4) is linear in the residual sum of squares, while for unknown σ , (3) is linear in the log of the residual sum of squares. The estimated parameters $\hat{\lambda}$ and $\hat{\beta}$ will change accordingly. If the fixed scalar σ and the estimated $\hat{\sigma}$ are close to each other, the differences between the estimates will be quite small. Least squares estimation uses (4) with the appropriate N_p . Neither $\hat{\lambda}$ or $\hat{\beta}$ changes compared to maximum likelihood, although N rather than N_p is used in (4), as the first term in (4) is constant relative to λ . In contrast, when using the restricted maximum likelihood, the profiled likelihood is modified by adding $-\frac{1}{2} \sum_{i=1}^M \log |(\mathbf{X}^*)^T \mathbf{X}^*|$ to (4), which will change the dependency from λ and the parameters $\hat{\lambda}$ and $\hat{\beta}$ as well. For general and mixed linear models, the **nlme** package uses, by default, the restricted maximum likelihood method.

If the co-variance matrix only depends on the variance σ^2 because Λ is a constant matrix, the equations (3) and (4) will attain their maximum at the same values of β . Thus, the estimates of the fixed parameters β will be identical regardless of whether σ is kept fixed or estimated and only the standard errors of the fixed coefficients will change.

The linear mixed model

The extension of the general linear model to a mixed linear model complicates the likelihood and subsequently, the estimation of its parameters. The model is known as the Laird-Ware model (Laird and Ware, 1982), but the original proposal for solving the equations is credited to Harville (Harville, 1977). If M independent groups and Q hierarchical levels are above the basic null level, a general mixed model complicates the presentation of its expression. As an illustration, a linear mixed model with one hierarchical random level only is presented here:

$$y_i = \mathbf{X}_i\beta + \mathbf{Z}_ib_i + \varepsilon_i \quad b_i \sim N(0, \Psi) \quad \varepsilon_i \sim N(0, \sigma^2\Lambda_i) \quad (5)$$

The matrices \mathbf{X}_i and \mathbf{Z}_i are the design matrices for p fixed parameters and q_1 random effects, respectively. The fixed parameters and random effects are vectors of length p and q_1 respectively, while y is a column vector of length n_i . A model with Q random levels would have $[q_j, j = 1, \dots, Q]$ random effects in total. It is assumed that b_i and ε_i are independently normally distributed with a mean of zero. The covariance matrix between each of the separate observations - sometimes called the basic random level - is similar to the one in (1), allowing different dependencies within this basic random level; e.g., correlations over time of observations grouped by subjects. The covariance matrix Ψ is the same for all groups. The latter may depend on a small number of parameters θ similar to the covariance matrices Λ_i , which depend on parameters λ . The parameters θ are called random coefficients. In (5) the expectation of b_i is assumed to be zero. The latter presents no restriction, as the expectations can be part of the fixed model parameters β_i . The expression (5) is redundant in the sense that the matrices Ψ and Λ_i could have been combined, but the previous formulation is mathematically more convenient. Likewise, in equation (2), an appropriate transformation of both sides of the expression will change the model into a standard linear mixed model with independent errors conditionally on the unknown parameters λ and θ .

A particularity of model (5) is that the vectors b_i are neither observed nor are parameters of the model itself. In analogy to Bayesian modelling, the effects b_i are integrated out of the formula (5). The resulting likelihood depends only on β , σ^2 , θ , and λ . The mathematical difficulty is that the random parameters θ are unknown and moreover the integration is multidimensional. Once the random parameters including σ^2 are estimated, these are plugged into the model to estimate the fixed parameters, similar to the general linear model. This approach is also known as empirical Bayesian estimation of a mixed model. In contrast to other approaches used in PROC MIXED and PROC GLIMMIX from SAS, the algorithm in `nlme` does not perform the integration numerically. In contrast, Pinheiro uses an analytical approach and integrates 'manually', which results in a solution exclusively in terms of ordinary matrix algebra, as in the general linear model (Pinheiro and Bates, 2001). In theory, this algorithm should be faster and more stable than that of PROC MIXED or PROC GLIMMIX from statistical analysis software (SAS). Stability in the sense of convergence depends on the model itself and is changeable. The drawback of the algorithm used by Pinheiro is that its derivation depends completely on the use of normal distributions, while the SAS algorithm can more easily be adapted for other distributions (such as binomial or Poisson). Both of the SAS procedures allow a full Bayesian solution by Monte Carlo Markov Chain (MCMC). Although b_i are not formally part of the model, these may be estimated and are called the random residuals of the model for grouped observations (e.g., subjects) and might be used to detect outlying groups.

In the following example, the (minus) log-likelihood of the standard mixed model for one hierarchical level ($Q = 1$) is compared to its equivalent of the model with known σ . For clarity of notation, the superscript $*$ is excluded from all design matrices, but one must bear in mind that the latter matrices now depend on the random parameters $\hat{\theta}$ as well as $\hat{\lambda}$. Thus changing the estimates of the latter parameters will change all the others as well. Similar to the general linear model, the use of the normal or restricted maximum likelihood will change the parameters as well, but for mixed models also the fixed parameters. As noted in Pinheiro and Bates (Pinheiro and Bates, 2001), the profiled (minus) log-likelihood is formulated as:

$$\begin{aligned} \text{lik}(\lambda, \theta | y, \sigma) = & \frac{N_p}{2} \left[\log(N_p) - \log(2\pi) - 1 - \sum_{i=1}^M \log \left(\|\mathbf{c}_{-1}^i\|^2 \right) \right] \\ & + \sum_{i=1}^M \log \left| \frac{\Delta}{\mathbf{R}_i} \right| \end{aligned}$$

The vectors \mathbf{c}_{-1}^i and the matrices \mathbf{R}_i are the result of a so-called QR decomposition of the design matrices \mathbf{X}_i . See Appendix 2 for details of the properties of QR decomposition. The matrices Δ are relative precision factors, which may actually be a vector of scalars. In Pinheiro and Bates (Pinheiro and Bates, 2001) the precision factors are defined as any matrix Δ that satisfies

$$\sigma^2 \Psi^{-1} = \Delta^T \Delta \quad (6)$$

Analogous to (4) the profiled (minus) log-likelihood can be expressed as:

$$l_{lik}(\lambda, \theta | y, \sigma) = \frac{N_p}{2} \left[-\log(2\pi) - \log(\sigma^2) \right] - \frac{\sum_{i=1}^M \|\mathbf{c}_{-1}^i\|^2}{2\sigma^2} + \sum_{i=1}^M \log \left| \frac{\Delta}{\mathbf{R}_i} \right| \quad (7)$$

The profiled log-likelihood with fixed σ depends linearly on a term that is similar to the residual sum of squares, while the equation of the usual mixed model depends on the log of the latter quantity. Again, N_p with the maximum likelihood of estimation equals N , while for restricted maximum likelihood (REML), $N_p = N - p$. In case of REML-estimation equation (7) is augmented

by: $\log(|\mathbf{R}_{00}|) = 0.5 \log \left(\left| \sum_{i=1}^M \mathbf{X}_i^T \Sigma_i^{-1} \mathbf{X}_i \right| \right)$ where the matrix \mathbf{R}_{00} is part of the QR decomposition.

Maximizing the augmented equation causes different estimates of the random parameters especially as compared to those from the maximum likelihood method (ML). The method of estimation by lme is, by default, the restricted maximum likelihood.

The non-linear mixed model

The non-linear mixed model is not only more complicated, its model structure is also different. In a linear model, the expectation of response is directly modelled as a linear combination of the row vectors of the design matrix \mathbf{X} . The fixed coefficients of the model define these linear combinations, while the random coefficients weight the latter for each group of the hierarchical level. By contrast, in the non-linear models, the expectation of the response is a non-linear function and each of its arguments is given by a mixed model. The non-linear model is described in detail by Davidian and Giltinan (Davidian and Giltinan, 1995) and its algorithms in Bates and Chambers (Bates and Chambers, 1988), (Bates and Watts, 1992), (Lindstrom and Bates, 1990) and Pinheiro and Bates (Pinheiro and Bates, 2001). The inclusion of the random parameters considerably complicates maximizing the log likelihood. In this section, the model and the profiled likelihood are outlined. For details and alternative algorithms for the same model, see Pinheiro and Bates (Pinheiro and Bates, 2001). As a point of departure, a single hierarchical level, repeated measures, non-linear model is explained next. For each of $i = 1, \dots, M$ groups and $j = 1, \dots, n_{ij}$ observations within a group, we have:

$$\begin{aligned} y_{ij} &= f(\phi_{ij}, v_{ij}) + \varepsilon_{ij} & \varepsilon_{ij} &\sim N(0, \sigma^2 \mathbf{I}) \\ \phi_{ij} &= \mathbf{A}_{ij}\beta + \mathbf{B}_{ij}b_i & b_i &\sim N(0, \Psi) \end{aligned}$$

Again, the vector β are fixed parameters and the vectors b_i are the random effects. The matrices \mathbf{A}_{ij} and \mathbf{B}_{ij} play a similar role as \mathbf{X}_i and \mathbf{Z}_i , respectively, in (5). The symbol f stands for a continuous non-linear differentiable function of one dimension, and v_{ij} is a vector of covariates. The model assumes that for each group (e.g., subjects), the arguments of the function may be different but its form is the same for all groups. The model can be straightforwardly extended for more levels of the hierarchy, or even by omitting the random terms as in gnls.

The nlme algorithm consists of alternating between a penalized non-linear least squares step (PNLS) and the linear mixed model step (NLM), which is similar and explained in the linear mixed model section. Therefore, only the consequences of fixing σ in the PNLS step have to be considered. The PNLS step minimizes:

$$\sum_{i=1}^M \left[\|y_i - f_i(\beta, b_i)\|^2 + \|\Delta b_i\|^2 \right] \quad (8)$$

The minimum is found by keeping in each step the precision factor (6) and the random coefficients b_i fixed. A convenient technique is to transform (8) into an ordinary least squares problem by augmenting the response for each group with a vector of 0's and the vector of function values by the column vector Δb_i . The sum of squares (8) is then identical to

$\sum_{i=1}^M \left[\|\tilde{y}_i - \tilde{f}_i(\beta, b_i)\|^2 \right]$ with \tilde{y}_i and \tilde{f}_i are the augmented vectors. Thus in the PNLS step, the sum of squares is conditional on the value of σ and either estimation in the mixed model step or substitution with a prior known value will yield the same solution. Only the likelihood function in the NLM step is affected by the choice of substitution of a known σ or one that is estimated. For the non-linear mixed model algorithm within each PNLS step, the same correction can be used as in the linear mixed model. A strong point of the algorithm devised by Pinheiro for the (non) linear mixed model is the strategy of getting solutions, not only by the classical root finding of the derivatives of the likelihood, but also by means of the expectation-maximization (EM)-algorithm. The latter does not require any derivatives. The EM algorithm is powerful because starting values of the parameters may be far from optimal

Table 1: Examples used for comparison of R and S-Plus output with and without fixed σ . References are to Pinheiro and Bates (Pinheiro and Bates, 2001)

Model type	nlme Function	Data set	Model name	Model function	Reference
Generalized linear	gls	Orthodont	fm1Orth.gls		p. 251
Non-linear	gnls	Dialyzer	fm1Dial.gnls	SSasympOff	p. 402
Mixed linear	lme	Orthodont	fm1Orth.lme		p. 147
	lme	Orthodont	fm3Orth.lme		p. 177
	lme	Wafer	fm1Wafer.lme		p. 172
Mixed non-linear	nlme	Theophylline	fm1Theo.nlme	SSfol	p. 363, 516
	nlme	Theophylline	fm2Theo.nlme	SSfol	p. 364, 516
	nlme	Theophylline	fm3Theo.nlme	SSfol	p. 365, 516
	nlme	Orange	fm1Oran.nlme	SSlogis	p. 358, 519

values. As on the other hand convergence slows when actually approaching the optimal values, the root finding takes over. Again, this property makes the algorithm quite robust in theory, though in practice, different starting values may cause differences in estimated values. Another difference between the algorithm of R and SAS is that in the former, the likelihood is re-parameterized using $\log(\sigma)$ rather than in σ , thus preventing the algorithm to crash if $\hat{\sigma}$ is close to zero. This problem is encountered frequently for over-parameterized models when using SAS. The **nlme** package contains a function `gnls` for the general non-linear model analogous to the general linear model function `gls`. It uses basically the same algorithm as `nlme` by stripping the random component.

Implementation and examples

From the viewpoint of the user, the only change in a function call of `gls`, is an extra argument in the control functions `glsControl`, `gnlsControl`, `lmeControl` and `nlmeControl` respectively. The option is implemented exactly as the corresponding control functions in S-Plus; for example, using a simple model and data from the S-Plus and R libraries. The output object of the functions `lme`, `nlme`, `gls`, and `gnls` contained a logical flag as component to signal whether σ is either fixed or estimated. The flag is used in the utility functions `print`, `summary`, `anova`, and `intervals`. Table 1 gives an overview of the examples used to test the new feature. The former are taken from Pinheiro and Bates (Pinheiro and Bates, 2001). All examples ran both in R and S-Plus.

In the next subsections the output from package **nlme** of R and library `nlme` of S-Plus are compared. Striking differences are given in the discussion. All scripts, output and difference files- in HTML format- are made available to the editors of the journal. Not all models converged. They were either in both R and S-Plus or in R or S-Plus only. The package **nlme** has a choice for two different optimizing algorithms but only the default has been used as this is the only one available in S-Plus. Thus, non-convergence in R might be caused by the use of the default algorithm. However, this has not been tested systematically. The same starting values have been used in both R and S-Plus, except in the examples with the self-starting functions in **nlme**. For the latter, starting values had to be supplied. The same examples for estimated σ have been run both for the original **nlme** version and the updated version. As the output was completely the same, eventual differences between S-Plus and R in case of estimated σ cannot be attributed to the update of the package. We have chosen to compare the R-output and the actual output of S-Plus 8.2 rather than the published one (Pinheiro and Bates, 2001), as in some cases, differences in the case of estimated σ were found between the latter and the actual version of S-Plus 8.2. At the time of its publication, fixed σ was not yet implemented.

General linear and general non-linear models

As an example, the code for the general linear and general non-linear model are given below for fixed σ .

```
fm10rth.gls <- gls( distance ~ Sex * I( age - 11), Orthodont)
```

While the same model with a fixed $\sigma = 2$ is declared as:

```
fm10rth_fix.gls <- gls( distance ~ Sex * I( age - 11), Orthodont,
                      control = glsControl(sigma = 2)).
```

The default estimates σ and the argument `sigma` to `glsControl` or `gnlsControl` (below) can be omitted or set to 0 or NULL. For both estimated and fixed σ , the results from the R and S-Plus estimates of the

coefficients were identical to the third decimal using either the estimation method of the maximum likelihood (ML) or the restricted maximum likelihood method (REML). However, for both the default of estimated σ and fixed σ , the equivalence of the correlation structure and the confidence limits of the coefficients might be limited only to the second decimal. An example of the general non-linear model is scripted as:

```
fm1Dial.gnls <- gnls( rate ~ SSasymOff( pressure, Asym, lrc, c0), Dialyzer,
  params = list( Asym + lrc ~ QB, c0 ~ 1),
  start = c(53.6, 8.6, 0.51, -0.26, 0.225),
  control = gnlsControl(sigma = NULL))
```

For fixed σ the argument `sigma` to the `gnlsControl` function should be changed to 1. The function `gnls` can only produce a least squares solution and rarely uses the C-library. Note that although the function `SSasymOff` is a self-starting function, starting values had to be supplied in R, but not in S-Plus.

Mixed linear models

The `gls` example from above was also fitted as a mixed model with the appropriate changes in the script:

```
fm10rth.lme <- lme( distance ~ I(age-11), Orthodont)
and
fm10rth_fix.lme <- lme( distance ~ I(age-11), Orthodont,
  control = lmeControl(sigma = 1))
```

Similar results as for the `gls` model were found for the ML method. However, both in R and S-Plus the REML estimation method for both estimated and fixed σ stopped due to a convergence error. The latter is remarkable as the estimates for the REML method are given in Pinheiro and Bates (Pinheiro and Bates, 2001)). The third example uses a covariance structure and is given as:

```
fm30rth.lme <- lme( distance ~ Sex * I( age - 11), Orthodont,
  weights = varIdent(form = ~ 1 | Sex))
```

While the same model with a fixed $\sigma = 2$ is declared as:

```
fm30rth_fix.lme <- lme( distance ~ Sex * I( age - 11), Orthodont,
  weights = varIdent(form = ~ 1 | Sex),
  control = list(sigma = 2)).
```

The ML method of estimation for both estimated and fixed σ were equal in the third digit, with the exception of the random effects and the interval estimates, in both cases of estimated and fixed σ . Differences were sometimes found to the second decimal. The same applies to the method of REML estimation. Finally, the corresponding R code for the linear mixed model example:

```
fm1Wafer.lme <- lme( current ~ voltage + I( voltage^2 ), Wafer,
  random = list( Wafer = pdDiag( ~ voltage + I( voltage^2 ) ),
  Site = pdDiag( ~ voltage + I( voltage^2 ) ) ),
  control = lmeControl( sigma = NULL))
```

For fixed σ , one has to substitute NULL by 1 in the argument `sigma` of `lmeControl`. Similar results are found in the two previous examples. But notable differences in the confidence limits of the random effects were seen when using the REML method with fixed σ , although the point estimates of the coefficients themselves were the same. For the interpretation of the correlation between the random effects, this might have consequences, as in R the latter did not contain 0 in the interval as opposed to the result from S-Plus 8.2. The latter could be caused by the difference in the degrees of freedom of the t-distribution, as our implementation of these might have changed if σ is kept fixed.

Non-linear mixed models

The `nlme` function for solving a non-linear problem defaults to the ML method as opposed to `lme`. During testing it was discovered that the option of the restricted maximum produced exactly the same estimates as for ML, with and without estimation of σ . In our opinion, this is not correct, but as far as we know, has never been reported. The example codes as:

```
fm1Theo.nlme<-nlme(conc ~ SSfol(Dose, Time, lKe, lKa, lCl),
  Theoph, fixed = lKe + lKa + lCl ~ 1,
  start=c( -2.4, 0.45, -3.2),
  control = nlmeControl( sigma = NULL))
```

Fitting σ requires the same change as in the above examples. The ML estimates using estimated σ were the same for R and S-Plus to the third digit. However, for fixed σ the example in S-Plus did not converge, and no comparison can be made. For the REML method, neither of the estimations (free or fixed) converged. R required starting values, even if SSfol was supposed to be self-starting. The next example from Table 1 is an extension of the above and the code is:

```
fm2Theo.nlme<-nlme(conc ~ SSfol(Dose, Time, lKe, lKa, lCl),
  Theoph, fixed = lKe + lKa + lCl ~ 1,
  start=c( -2.4, 0.45, -3.2),
  control = nlmeControl( sigma = NULL),
  random = pdDiag( lKe + lKa + lCl ~ 1))
```

The estimates of the fixed coefficients are equal to the second or third place in all cases. In contrast differences between the random parameters for ML, REML, and estimated and fixed σ are larger. Also, the approximate variance-covariance matrix of the random effects is nearly singular, which causes the 95% lower and upper limits for lKe to be near infinity. In S-Plus, the covariance matrix of the REML random estimates was not estimated at all due to singularity of the Hessian matrix. This may indicate that the parametrization of model is not correct, probably for parameter lKe. The next example uses the same data set but omits lKe from the random parameter list:

```
fm2Theo.nlme<-nlme(conc ~ SSfol(Dose, Time, lKe, lKa, lCl),
  Theoph, fixed = lKe + lKa + lCl ~ 1,
  start=c( -2.4, 0.45, -3.2),
  control = nlmeControl( sigma = NULL),
  random = pdDiag( lKa + lCl ~ 1))
```

Now the outcomes of all the models are in better agreement, and the covariance of the random effects is available for S-Plus. There are still differences from the third digit onwards for the random effects confidence limits, especially for the REML method.

Note that for all of the above models we have changed the degrees of freedom of the log likelihood in case σ is held fixed, which coincides with the output of S-Plus. This may cause differences in the confidence limits especially when the degrees of freedom of the t-distribution is small.

A case study illustrating the application of a non-linear mixed model with a binomial response

A summary clinical outcome dataset including clinical outcome data from 14 randomized controlled trials of 4 drugs was extracted from the Quantify RA clinical database developed by Quantitative Solutions (drug and trial names blinded). Three endpoints were included describing the proportion of patients with ACR20, ACR50, and ACR70 responses. These endpoints are based on the American College of Rheumatology (ACR) criteria used to assess improvement in tender or swollen joint counts and improvement in three of the following five parameters: acute phase reactant (such as sedimentation rate), patient and physician assessment, pain scale, and a disability/functional questionnaire. For ACR20 response, a 20 percent improvement in tender or swollen joint counts as well as 20 percent improvement in three of the other five criteria is required. Similarly, for ACR50 and ACR70, a 50 and 70 percent improvement is required, respectively. Outcomes were available up to 6 months (24 weeks) after the start of treatment with one of the drugs, up to 12 weeks with the two other drugs, while the last drug was only observed in a 4-week trial. The analysis dataset was analyzed using two non-linear regression models and a mixed effects model with random effects on the level of trial assuming binomial distributed data. The probability of ACR20/50/70 response is confined to the interval between zero and one. In this example, the logit transformation was used as a link function between the probability of response p and a hidden model variable X :

$$p = \frac{e^X}{1 + e^X}$$

Table 2: Parameter estimates with uncertainty based on a gnls model with constant placebo. Results for estimated and fixed σ in left and right column respectively.

Parameter	Estimate (%RSE) Estimated σ	Estimate (%RSE) Fixed $\sigma = 1$
E_0	-1.06 (10%)	-1.11 (4%)
α_{ACR50}	-1.05 (5%)	-1.07 (3%)
α_{ACR70}	-1.97 (3%)	-2.00 (2%)
E_{MAX}	2.02 (10%)	2.06 (4%)
$\log ED_{50,drug1}$	2.22 (63%)	2.14 (31%)
$\log ED_{50,drug2}$	1.75 (20%)	1.68 (9%)
$\log ED_{50,drug3}$	4.11 (16%)	4.04 (8%)
$\log ED_{50,drug4}$	1.22 (36%)	1.22 (15%)
Res. error	1.924	1 (fixed)

A convenient property of the logit transformation is that the log odds ratio (OR) of two probabilities translates into the difference between the corresponding hidden model variables:

$$\log OR = \log \left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \right) = X_1 - X_2$$

If X is described by a linear model, correction for any predictor can be done by subtraction. Specifically, if the probability of response is modelled as the inverse logit of the sum of a placebo (E_0) and a drug effect (g), the log OR of the treated versus the untreated response is simply given by function g ; (see 11). The logit transformed observed proportion of ACR20/50/70 responders was described as the sum of an unstructured placebo response ($E_{0,it}$) and drug-dependent drug effects described by EMAX models with common E_{MAX} and drug-specific ED_{50} estimates.

$$N_{response,kijt} \sim \text{binomial} \left(P(response)_{kijt}, N_{kijt} \right) \quad (9)$$

$$P(response)_{kijt} = \text{logit}^{-1} \left(E_{0,it} + g_k \left(Drug_{ij}, Dose_{ij} \right) + \alpha_k \right) \quad (10)$$

$$g_k \left(Drug_{ij}, Dose_{ij} \right) = \frac{E_{MAX} \cdot Dose_{ij}}{ED_{50,Drug_{ij}} + Dose_{ij}} \quad (11)$$

Where $N_{response,kijt}$ is the number of subjects in treatment arm j of trial i with ACR_k response at the t_{th} visit out of N_{kijt} subjects. The coefficient α_k estimated for ACR50 and ACR70 describes the constant log OR of these endpoints relative to ACR20. Three variations of the model described in (11) were investigated using fixed and estimated residual standard error. Model 1 was a gnls model including a constant placebo model across trials and visits; i.e. $E_{0,it} = E_0$. Model 2 was a gnls model including an unstructured placebo model estimating a non-parametric placebo response at each visit in each trial; $E_{0,it}$ as shown in (11). Model 3 was a non-linear mixed effects model including an unstructured placebo model with uncorrelated random effects on E_{MAX} and α_k on the level of trial. All three models included a compound symmetry model correlating outcomes from different endpoints and visits within a trial. An overview of the parameter estimates for the three models are in Tables 2, 3, and 4 respectively.

It is interesting to note that comparison with anova showed significant differences between free estimated and fixed σ for models GNLS models 1 and 2, whereas the same comparison for the nlme model 3, did not show a significant difference. The (lack of) differences appeared for the criteria AIC and BIC, as well for the likelihood ratio test. For the former criteria the usual difference of 2 was used as a threshold. See Tables 5, 6 and 7. Comparison between models 1, 2 and 3 are not shown but from the tables it is clear that these models have an increasingly lower AIC and BIC, and better log likelihood ratio's either for free or fixed σ .

Conclusions and discussion

The estimates of the fixed coefficients of S-Plus and R are often identical to three significant digits, except in the case of over-parameterized models, which may cause non-convergence and inaccurate estimation of the covariance matrix of the random parameters. Differences might occur even for the default model with estimation of σ . For the non-linear mixed models differences between R and

Table 3: Parameter estimates with uncertainty based on a gnls model with unstructured placebo. Results for estimated and fixed σ in left and right column respectively.

Parameter	Estimate (%RSE) Estimated σ	Estimate (%RSE) Fixed $\sigma = 1$
Mean* E_0	-1.16	-1.22
α_{ACR50}	-1.08 (3%)	-1.09 (3%)
α_{ACR70}	-2.05 (2%)	-2.06 (2%)
E_{MAX}	2.24 (8%)	2.26 (5%)
$\log ED_{50,drug1}$	2.87 (69%)	2.80 (50%)
$\log ED_{50,drug2}$	1.94 (13%)	1.92 (9%)
$\log ED_{50,drug3}$	4.47 (21%)	4.42 (15%)
$\log ED_{50,drug4}$	1.66 (28%)	1.44 (22%)
Res. error	1.313	1 (fixed)
E_0^* : Mean of 34 unstructured placebo parameters		

Table 4: Parameter estimates with uncertainty based on a nlme model with unstructured placebo. Results for estimated and fixed σ in left and right column respectively.

Parameter	Estimate (%RSE) Estimated σ	Estimate (%RSE) Fixed $\sigma = 1$
Mean* E_0	-1.28	-1.26
α_{ACR50}	-1.08 (8%)	-1.08 (8%)
α_{ACR70}	-2.08 (5%)	-2.08 (5%)
E_{MAX}	2.34 (11%)	2.33 (11%)
$\log ED_{50,drug1}$	2.97 (50%)	2.99 (54%)
$\log ED_{50,drug2}$	1.78 (10%)	1.76 (11%)
$\log ED_{50,drug3}$	4.66 (16%)	4.67 (17%)
$\log ED_{50,drug4}$	1.59 (33%)	1.71 (32%)
$sd(\alpha_{ACR50})$	0.260	0.257
$sd(\alpha_{ACR70})$	0.351	0.344
$sd(E_{MAX})$	0.761	0.745
Res. error	0.938	1 (fixed)
E_0^* : Mean of 34 unstructured placebo parameters		

Table 5: Analysis of variance table for comparison of gnls model 1 between free and fixed σ

Model	df	AIC	BIC	logLik	L.Ratio	p-value
model 1 free σ	10	-740.2	-700.1	380.1		
model 1 fix σ	9	-271.1	-235.0	144.5	471.2	<.0001

Table 6: Analysis of variance table for comparison of gnls model 2 between free and fixed σ

Model	df	AIC	BIC	logLik	L.Ratio	p-value
model 2 free σ	43	-1100.5	-928.0	593.3		
model 2 fix σ	42	-1076.8	-908.4	580.4	25.7	<.0001

Table 7: Analysis of variance table for comparison of nlme model 3 between free and fixed σ

Model	df	AIC	BIC	logLik	L.Ratio	p-value
model 3 free σ	46	-1215.8	-1031.3	653.9		
model 3 fix σ	45	-1215.8	-1035.3	652.9	2.0	0.16

S-Plus were more likely to occur when using REML estimation. Sometimes REML estimation was not possible in either R or S-Plus due to lack of degrees of freedom.

There are some differences between R and S-Plus in the code of the functions. As the code of the C-functions in S-Plus is not public, differences in the code itself could not be checked. The R functions, in particular the interaction between the R-code and the C-functions were fragile in the sense that input was often changed within a function and then used as output. Thus, a small change in code in one part could have unexpected results in other parts of the output.

The code of the **nlme** package is far from modular; e.g., parts of both user-accessible and hidden functions of **lme** were used in **nlme** or **gls** causing lack of transparency. For example, within the function `print.summary.gls`, at some point the function `print.summary.lme` was used, which caused difficulties with passing the right degrees of freedom of the likelihood. Different definitions of the likelihood functions coded in C were used in the R-code of the same parent functions, amongst others in **lme**. Another critical point is that utilities as `print`, `summary` or `print.summary` functions tend to recalculate again some statistics such as the likelihood, AIC or BIC. For these reasons, debugging has been time consuming and given the complexity of the **nlme** package the number of tests provided by CRAN is simply too small. Thus, the warning ‘the use of free software comes with no guarantee’ in the code of each of the functions must be taken in earnest.

A general drawback of the algorithm used in **nlme** is that it is completely geared to the use of the normal distribution. The latter makes it difficult - though not impossible - to adapt the algorithm for other distributions such as binomial or Poisson. The CRAN-library already contains the package **nlme4** allowing non-normal distributions, but at this time lacks the feature of the use of correlated observations. Perhaps a future update will be available that contains an option for fixing the scale parameter σ .

In the **nlme** library of S-Plus the degrees of freedom for residual error and consequently, the t-tests, are not properly adapted for fixed σ . However, in the function `anova`, the degrees of freedom has been changed in S-Plus to allow comparison between models with and without fixed σ . This correction affects both AIC and BIC. For example, AIC will always decrease by 2 while BIC will decrease by $2 \times \log(N)$. Apart from that, the value of the profiled log likelihood for the REML estimation should change, because the degrees of freedom is part of the latter function. For reasons of compatibility with S-Plus, the changes in the degrees of freedom in the t-tests have not been implemented in **nlme** of R.

When comparing the output of **nlme** between R and S-Plus, a flaw became apparent in the output of `anova` if used with a single object. The F-tests for the (grouped) parameters differed between both programs. The F-values from R can be made equivalent to those from S-Plus by multiplying the F-values by $\frac{N}{N-p}$. The same applies to the standard errors of the fixed estimates. There the factor should be $\sqrt{\frac{N}{N-p}}$. We believe that the output from S-Plus is correct considering that a t-test with one degree of freedom is equivalent to an F-test with one degree in the numerator in which case, the latter is equivalent to the square of the former. The finding that, in the function **nlme** in R, that there is no difference between ML and REML estimation is remarkable and may be due to an error in the code or to a compelling reason by the maintainer of the package.

From a statistical point of view, a warning should be issued on the use of the likelihood ratio test when comparing two models; one with a fixed and one with an estimated σ . In fact, the usual F-test is not appropriate. The reason is that the log of both numerator and denominator have an approximate Chi-square distribution up to the log of the true value of σ , which is unknown if σ is estimated. The difference in the log-ratio between the known and estimated value of σ does not go away, which causes a bias and invalidates the F-distribution of the log ratio. A safe way to test whether the a priori value is right, is to use AIC or BIC using the recommendations by Raftery (Raftery, 1980). As Raftery states, difference between AICs of 2 or less is ‘hardly worth mentioning’. As previously mentioned, an increase of one degree of freedom of the model already causes a decrease of 2 in the AIC, even if the likelihood does not change. One could then be tempted to prefer the model with more degrees of freedom although Raftery seems not to recommend this. The influence on the BIC is even stronger, as its decrease will be at least $2 \times \log(\sigma)$. With the developed patches, we have successfully back-ported the ability to fix the residual standard error in the **nlme** package from S-Plus to R. Following review, the patches have been accepted into the official version **nlme** (available through CRAN). This will allow users to correctly estimate models where the residual standard error is known, using standard and publicly available open-source tools.

Acknowledgements

Special thanks go to Martin Maechler for critically looking at the added code, running the CRAN tests independently and fine tuning our coding. We are greatly indebted to Mrs. Flowers Lovern for scrutinizing the English text, both on syntax and idiom.

De-correlation of the normal model

Starting from a normal model with correlated errors, it is shown below that this model can be converted into a model with independent errors and equal variance. Let Λ be a positive-definitive matrix, then a non-singular square root of that matrix exists, $\Lambda = \left(\Lambda^{1/2}\right)^T \Lambda^{1/2}$. Similarly, for the inverse of the matrix. The original response vector \mathbf{y} and the errors ϵ are now both multiplied by the inverse of the square root matrix, and are $\mathbf{y}^* = \left(\Lambda^{-1/2}\right)^T \mathbf{y}$ and $\epsilon^* = \left(\Lambda^{-1/2}\right)^T \epsilon$ respectively. The transformed design matrix is $\mathbf{X}^* = \left(\Lambda^{-1/2}\right)^T \mathbf{X}$. The covariance matrix of the transformed observations is then $V(\epsilon^*) = \sigma^2 \left(\Lambda^{-1/2}\right)^T \Lambda \Lambda^{-1/2} = \sigma^2 \mathbf{I}$, and the fixed coefficients β of the model can now be solved with the usual least squares method.

QR decomposition

A $N \times p$ matrix \mathbf{X} of rank p may be decomposed in an orthogonal matrix \mathbf{Q} of dimension $n \times n$ and a $p \times p$ upper triangular matrix \mathbf{R} . More specifically

$$\mathbf{X} = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_t \mathbf{R}$$

The matrix \mathbf{Q}_t is the truncated matrix \mathbf{Q} consisting of the first p columns of the latter matrix. An orthogonal matrix is defined as $(\mathbf{Q}_t)^T \mathbf{Q}_t = \mathbf{I}$. The decomposition is performed in the model (5) for all matrices to simplify the actual numerical computations as well as the representation of the estimation equations. The reason is that multiplication of a vector y augmented with p 0's conserves the Euclidian norm of that vector. So $\|(\mathbf{Q}_t)^T y\|^2 = \|y\|^2$. Similarly, the vector of residuals of a regression model $\|y - \beta \mathbf{X}\|^2$ is conserved and equals $\|\mathbf{c}_1 - \mathbf{R} \beta\|^2 + \|\mathbf{c}_2\|^2$. The vector $\mathbf{c}^T = (\mathbf{c}_1, \mathbf{c}_2)$ with length p and $n - p$ and equals $(\mathbf{Q}_t)^T \mathbf{y}$. Maximizing the corresponding log likelihood yields the least squares estimator $\hat{\beta} = \mathbf{R}^{-1} \mathbf{c}_1$ and the residual sum of squares equals $\|\mathbf{c}_2\|^2$.

Bibliography

- H. Akaike. Likelihood and the Bayes procedure. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*. University Press, Valencia, Spain, 1980. [p239]
- D. M. Bates and J. M. Chambers. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, New York, U.S.A., 1988. [p243]
- D. M. Bates and D. G. Watts. Nonlinear models. In C. J.M. and H. T.J., editors, *Statistical Models in S*, pages 421–454. Wadsworth and Brooks, Pacific Grove, U.S.A., 1992. ISBN 0 534 16764 0. [p243]
- S. Beal and L. Sheiner. The NONMEM system. *The American Statistician*, 34:118–119, 1988. [p239]
- M. Davidian and D. M. Giltinan. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London, U.K., 1995. ISBN 0 412 98341 9. [p239, 240, 243]
- D. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72:320–340, 1977. [p242]
- N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38:963–974, 1982. [p239, 240, 242]
- M. J. Lindstrom and D. M. Bates. Nonlinear mixed-effects models for repeated measures data. *Biometrics*, 46:673–687, 1990. [p243]
- R. C. L. Littell, G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. *SAS for Mixed Models, Second Edition*. SAS Institute, 2006. ISBN 1590475003. [p239]
- J. C. Pinheiro and D. M. Bates. *Mixed Models in S and S-Plus*. Springer-Verlag, New York, U.S.A., 2001. ISBN 0 387 98957 9. [p239, 240, 241, 242, 243, 244, 245]
- A. E. Raftery. Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, Spiegelhalter, and D.J., editors, *Markov Chain Monte Carlo In Practice*, pages 163–186. Chapman and Hall, London, U.K., 1980. ISBN 0 412 05551 1. [p249]

- G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978. [p239]
- S. N. Wood. *Generalized Additive Models, An Introduction with R*. Chapman and Hall, London, U.K., 2006. ISBN 1 58488 474 6. [p240]

S.H. Heisterkamp
BioStat-Plus B.V.
Mariaplaats 32E, 3511LL Utrecht
the Netherlands
info@biostat-plus.nl

E. van Willigen
Van Willigen ICT-consultancy
Kempke 19, 5672PL Nuenen
the Netherlands
e.vanwilligen@onsnet.nu

P.M Diderichsen
Quantitative Solutions B.V.
Parkstraat 1b, 4818SJ Breda
the Netherlands
pmdiderichsen@wequantify.com

J. Maringwa
Quantitative Solutions B.V.
Parkstraat 1b, 4818SJ Breda
the Netherlands
jmaringwa@wequantify.com