Model Building for Nonlinear Mixed Effects Models

José C. Pinheiro
Douglas M. Bates
Department of Statistics

Mary J. Lindstrom
Department of Biostatistics

Technical Report #931
August, 1994
Model Building in Nonlinear Mixed Effects Models

José C. Pinheiro, Douglas M. Bates, and Mary J. Lindstrom
Department of Statistics         Dept. of Biostatistics
University of Wisconsin — Madison

ABSTRACT

Nonlinear mixed effects models involve both fixed effects and random effects. Model building for nonlinear mixed effects is the process of determining the characteristics of both the fixed and the random effects so as to give an adequate but parsimonious model. We describe procedures based on information criterion statistics for comparing different structures of the random effects component. These include procedures for determining which parameters in the model should be mixed effects and which should be purely fixed effects, as well as procedures for modeling the dependence of parameters on cluster-specific covariates. These methods are illustrated using the nonlinear mixed effects methods and classes for S-plus and using data sets from a laboratory pharmacokinetic study and from a pharmacokinetics clinical study.

Keywords: Mixed effects, random effects, nonlinear models, maximum likelihood.

1. INTRODUCTION

Nonlinear mixed effects models have received a great deal of attention in the statistical literature in recent years because of the flexibility they offer in handling the unbalanced repeated measures data that can arise in different areas of investigation, such as pharmacokinetics, economics, and ecology.

Model building in mixed effects models involves questions that do not have a parallel in (fixed effects) linear and nonlinear models. Some of these questions are:

- determining which effects should have an associated random component and which should be purely fixed;
- using covariates to explain cluster-to-cluster parameter variability;
- using structured random effects variance-covariance matrices (e.g. diagonal matrices) to reduce the number of parameters in the model.

We consider here strategies for addressing these questions in the context of nonlinear mixed effects models, though most of the techniques described are also applicable to linear mixed effects models.

Any model building strategy is by nature iterative: a tentative model is initially fitted and modified to generate possibly better models (according to some goodness-of-fit criterion) and the process is repeated until no further improvements are possible. In comparing alternative models one must also analyze the residuals from the fit, checking for departures from the model's assumptions. It is also highly recommended that any model building analysis be done in conjunction with experts in the field of application of the model, to ensure the practical usefulness of the chosen model.

2. THE NONLINEAR MIXED EFFECTS MODEL

Several different nonlinear mixed effects models have been proposed in recent years (Sheiner and Beal, 1980; Mallet, Mentre, Steinner and Lokiec, 1988; Lindstrom and Bates, 1990; Vonesh and Carter, 1992; Davidian and Gallant, 1992; Wakefield, Smith, Racine-Poon and Gelfand, 1994). We consider here a slightly modified version of the model proposed in Lindstrom and Bates (1990). This model can be viewed as a hierarchical model that in some ways generalizes both the linear mixed effects model of Laird and Ware (1982) and the usual nonlinear model for independent data (Bates and Watts, 1988). In the first stage the jth observation on the ith cluster is modeled as

\[ y_{ij} = f(\phi_{ij}, x_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i \]  

(2.1)

where \( f \) is a nonlinear function of a cluster-specific parameter vector \( \phi_{ij} \) and the predictor vector \( x_{ij}, \epsilon_{ij} \) is a normally distributed noise term, \( M \) is the total number of clusters, and \( n_i \) is the number of observations in the \( i \)th cluster. In the second stage the cluster-specific parameter vector is modeled as

\[ \phi_{ij} = A_{ij} \beta + B_{ij} b_i, \quad b_i \sim N(0, \sigma^2 D), \]  

(2.2)

where \( \beta \) is a \( p \)-dimensional vector of fixed population parameters, \( b_i \) is a \( q \)-dimensional random effects vector associated with the \( i \)th cluster (not varying with \( j \)), \( A_{ij} \) and \( B_{ij} \) are design matrices for the fixed and random effects respectively, and \( \sigma^2 D \) is a...
(general) variance-covariance matrix. It is further assumed that observations made on different clusters are independent and that the \( \epsilon_i \) follow a \( \mathcal{N}(0, \sigma^2 A_i) \) distribution and are independent of the \( b_i \). We will restrict ourselves here to the case where \( A_i = I \).

The parameters in the model are estimated by either maximum likelihood, or restricted maximum likelihood, based on the marginal density of \( y \). In general this density does not have a closed-form expression when the model function \( f \) is nonlinear in \( b \), so different approximations have been proposed for estimating it. We adopt here the approximation suggested in Lindstrom and Bates (1990). Pinheiro and Bates (1994) analyze several approximations to the loglikelihood of nonlinear mixed effects models and conclude that Lindstrom and Bates’ approximation usually gives very accurate results.

3. Examples

We make extensive use of real data examples to illustrate the model building techniques discussed here. We now introduce the data sets that will be used throughout this paper.

3.1. Theophylline

This data set is courtesy of Dr. Robert A. Upton of the University of California, San Francisco. Theophylline was administered orally to 12 subjects whose serum concentrations were measured at 11 times over the next 25 hours. The data, presented in figure 1, is an example of a laboratory pharmacokinetic study characterized by many observations on a moderate number of individuals (clusters).

A common model for such data is a first order compartment model with absorption in a peripheral compartment

\[
C_t = \frac{D K_k a}{C_l (k_a - K)} \left[ \exp(-K t) - \exp(-k_a t) \right] \tag{3.1}
\]

where \( C_t \) is the observed concentration at time \( t \) (mg/L), \( t \) is the time (hr), \( D \) is the dose (mg/kg), \( C_l \) is the clearance (L/hr), \( K \) is the elimination rate constant (1/hr), and \( k_a \) is the absorption rate constant (1/hr). In order to ensure positivity of the rate constants and the clearance, the logarithms of these quantities can be used in (3.1), giving the reparametrized model

\[
C_t = \frac{D \exp(I K_a + I K - I C_l)}{\exp(I k_a) - \exp(I K)} \times \left\{ \exp[-\exp(I K) t] - \exp[-\exp(I k_a) t] \right\} \tag{3.2}
\]

where \( I C_l = \log(C_l) \), \( I k_a = \log(k_a) \), and \( I K = \log(K) \).

3.2. Quinidine

The second data set comes from a pharmacokinetics clinical study of the antiarrhythmic drug Quinidine. A total of 361 Quinidine concentration measurements were made on 136 hospitalized patients under varying dosage regimens. Additional data were collected on a set of nine covariates: age, height, weight, race, smoking status, ethanol abuse, congestive heart failure, creatinine clearance, and \( \alpha-1 \)-acid glycoprotein concentration. Some of these covariates varied for the same patient during the course of the study, while others remained constant. One of the main objectives of the study was to investigate relationships between the individual pharmacokinetic parameters and the covariates. A full description of the data can be found in Verme, Ludden, Clementi and Harris (1992). Statistical analyses of these data using alternative modeling approaches are given in Davidson and Gallant (1993) and Wakefield (1993).

The model that has been suggested for the Quinidine data is the one-compartment open model with first-order absorption. This model can be defined in a recursive way as follows.

Suppose that, at time \( t \), the patient receives at dose \( d_t \) and prior to that time the last dose was given at time \( t' \). The expected concentration, \( C_t \), and the apparent concentration in the absorption compartment, \( C_a_t \), are given by

\[
C_t = C_v \exp[-K (t - t')] + \frac{C_v k_a}{k_a - K} \left[ \exp[-K (t - t')] - \exp[-k_a (t - t')] \right] \times \left\{ \exp[-K (t - t')] - \exp[-k_a (t - t')] \right\} \tag{3.3}
\]

\[
C_a_t = C_v \exp[-k_a (t - t')] + \frac{d_t}{V}
\]

where \( V \) represents the apparent volume in distribution and \( k_a \) and \( K \) are respectively the absorption and the elimination rate constants.

When a patient receives the same dose \( d \) at regular time intervals \( \Delta \), model (3.3) converges to a steady state model, where
the expected concentrations are given by

\[ C_t = \frac{d k_a}{V (k_a - K)} \times \left[ \frac{1}{1 - \exp(-K\Delta)} - \frac{1}{1 - \exp(-k_a\Delta)} \right] \]

\[ C_a = \frac{d V}{V [1 - \exp(-k_a\Delta)]} \]

Patients considered to be in steady state conditions have concentrations modeled as above.

Finally, for a between-dosages time \( t \), the model for the expected concentration \( C_t \), given that the last dose was received at time \( t' \), is identical to (3.3).

Using the fact that the elimination rate constant \( K \) is equal to the ratio between the clearance (Cl) and the volume in distribution (V), we can reparametrize models (3.3) and (3.4) in terms of \( V, k_a \), and Cl.

In order to ensure that the estimates of \( V, k_a \), and Cl are positive, we can rewrite models (3.3) and (3.4) in terms of \( \log(V), \log(k_a) \), and \( \log(Cl) \).

The initial conditions for the recursive model are \( C_0 = 0 \) and \( C_{a0} = d_0/V \), with \( d_0 \) denoting the first dose received by the patient. It has been assumed throughout the model's definition that the bioavailability of the drug, i.e. the percentage of the administered dose that reaches the measurement compartment, is equal to one.

4. VARIANCE-COVARIANCE MODELING

In this section we consider the questions of determining which parameters in the model should have a random component and whether the scaled variance-covariance matrix of the random effects (\( D \)) can be structured in a simpler form (i.e. with fewer parameters than the unstructured form).

The first question that should be addressed in the analysis is choosing which parameters should be random effects and which purely fixed effects. Our approach is to fit different prospective models and compare nested models using likelihood ratio tests or some information criterion statistics, e.g. AIC (Sakamoto, Ishiguro and Kitagawa, 1986). One of the problems with this approach is deciding which way to construct the nesting; from smaller to larger models, or the other way around. Starting with a model where all parameters have associated random effects and then removing unnecessary terms is probably the best strategy, but may not be possible to implement if the model is badly overparametrized. In these cases the variance-covariance matrix of the random effects may become seriously ill-conditioned, making convergence difficult or impossible. The smaller to larger approach is an alternative in these cases, but has the disadvantage of the large number of models that may have to be fitted before the desired one is found. There is yet another important aspect that is overlooked by the model nesting approach: sometimes it is a linear combination of random effects being treated as fixed that gives the best model reduction.

The strategy we suggest for choosing the random effects to be included in the model is to start with all parameters as mixed effects, whenever no prior information about the random effects variance-covariance structure is available and convergence is possible. Then we examine the eigenvalues of the estimated \( D \) matrix, checking if one, or more, are close to zero. The associated eigenvalue(s) would then give an estimate of the linear combination of the parameters that could be taken as fixed. We used the Akaike information criterion to decide between alternative models, choosing the one with the smaller AIC.

Small eigenvalues may arise when the relative magnitude of the scales of the parameters in the model are quite different, without necessarily implying overparametrization. Therefore we suggest using a normalized, scale invariant version of the variance-covariance matrix. There are different ways of normalizing \( D \), the most common being the correlation matrix. This is not a particularly good choice in the present context, since all random effects would then have normalized variance equal to one and we would not be able to identify those with relatively small dispersion (which would be natural candidates to be dropped from the model). Whenever the \( A \) and \( B \) matrices in (2.2) are incidence-like matrices (i.e. with just one nonzero entry per row), a more convenient choice of normalization is the coefficient of variation (CV) matrix \( D_{CV} \) with

\[ [D_{CV}]_{ij} = \frac{[D]_{ij}}{\beta_k(i)\beta_k(j)} \]

where \( \beta_k \) represents the \( k \)th fixed effect and \( k(i), k(j) \) represent the indices of the fixed effects associated with the \( i \)th and \( j \)th random effects. When the nonzero elements of the \( i \)th row of \( A \) and \( B \) are equal to one, the \( i \)th diagonal element of \( D_{CV} \) is equal to the square of the coefficient of variation of \( \phi_k \). We note that, in the vast majority of real life applications of model (2.1), \( A \) and \( B \) will be incidence-like matrices.

To illustrate the use of this method we consider the examples described in section 3. We do not include any analyses of residuals here, but in all cases they did not indicate violations of the model's assumptions. All maximum likelihood calculations in the examples were done using the \texttt{n1me} function in S-plus (Pinheiro, Bates and Lindstrom, 1993).

4.1. Theophylline

The Theophylline data give an example where convergence is attained for the model in which all parameters are mixed effects. We refer to this model as model 1.

The AIC of model 1 is 124.03 and the MLE of the \( D_{CV} \) matrix has eigenvalues 4.324, 0.019, and \( 2.031 \times 10^{-7} \) indicating that the model is probably overparametrized. The eigenvector corresponding to the smallest eigenvalue, converted back to the
original scale and normalized is (0.464, 0.020, −0.886)\textsuperscript{T}, suggesting that the lCI random effect is approximately equal to twice the lK random effect. Recalling that the volume in distribution (V) is equal to the ratio between Cl and K, we see that lCI = 2lK implies that the ratio between V and K, that we will denote by R, is a fixed effect. The recommendation at this point would be to contact a pharmacologist and check the plausibility of this finding, as well as the interpretability of the parameter R. We will proceed the analysis here for the purpose of illustrating the use of the proposed model building techniques.

We reparametrized model (3.2) in terms of lka, lK, and lR = \log(R), letting only the first two parameters be mixed effects. The AIC of this reduced model, called model II, is 118.20, considerably smaller than the AIC of model I. The eigenvalues of the estimated D_{CV} matrix are 0.356 and 0.158 indicating that no further linear combinations of random effects can be eliminated from the model. In fact, if we remove either the lka or the lK random effect we get AIC values of respectively 203.865 and 200.135, both substantially worse than model II.

It is interesting to compare model II with the models obtained by considering each parameter at a time in model I as a fixed effect, to check if a more easily understood model could be used. The AIC of the models considering each of lCl, lka, and lK at a time as fixed effects are respectively 163.224, 194.189, and 125.446, all considerably larger than the AIC of model II. Note however that the elimination of the lK random effect from model I has a much smaller impact on the AIC value, than the elimination of either the lCl or the lka random effects. This suggests that if one is willing to correct the overparametrization problem by dropping one of the random effects from the model (and not a linear combination of them, as was done in model II), lK would be the natural choice.

The estimated correlation between the lK and the lka random effects in model II was −0.132, suggesting that the two random effects could be regarded as independent and a diagonal D used. The AIC of this model (III) is 116.388 indicating that it should be preferred over the previous models. No further reduction in the number of parameters in D could be obtained and we concluded that model III was the most adequate.

**4.2. Quinidine**

The Quinidine data provide an example where convergence cannot be attained for the model with all parameters as mixed effects. The data are characterized by few observations on many patients: for 46 patients there is only one observation of Quinidine concentration and for 32 patients only two. As a consequence, the optimization of the loglikelihood for the saturated model (called model I) becomes a very ill-conditioned numerical problem, with the optimizing algorithm alternating between equivalent solutions (in terms of the value of the loglikelihood) without ever converging.

Different strategies can be used to try to circumvent the non-convergence problem:

- try to achieve convergence using a diagonal D and examine the relative variability of the random effects, investigating the possibility of eliminating one, or more, of them from the model;
- force convergence (e.g. letting the algorithm run until a pre-established maximum number of iterations) and examine the corresponding D_{CV} matrix for rank deficiency;
- try to achieve convergence for models with a smaller number of random effects.

Convergence could not be achieved even for a diagonal D and so the second strategy was used here. We forced convergence after ten iterations of Lindstrom and Bates’ alternating algorithm. The AIC of this forced convergence fit was 344.74. The eigenvalues of the D_{CV} matrix were equal to 74.526, 0.032, and 1.363 \times 10^{-9}, suggesting that the model was overparametrized.

The eigenvector corresponding to the smallest eigenvalue, converted to the original scale of the random effects and normalized, was (0.097, 0.415, −0.905)\textsuperscript{T}, indicating that the lka random effect was about twice the lK random effect. In terms of the original parameters, that is equivalent to assume that R = k_a/V\textsuperscript{2} is a fixed effect. As in the Theophylline example, the recommendation at this point would be to consult a pharmacologist about the physical meaning, if any, of the parameter R. In this example, though, it seems that the sparse nature of the data is responsible for the convergence problems in general, and the rank deficiency observed for D_{CV} in particular. Therefore, by using a reparametrization of model I in which R was incorporated as a fixed effect, we would run the risk of overfitting low quality data. We decided to follow a more conservative approach trying to solve the overparametrization problem by removing each random effect at a time from model I.

Convergence was attained for the models in which each of lCl, lka, and lV at a time were treated as fixed. The corresponding AIC values were respectively 501.925, 341.782, and 365.409, indicating that the model in which lka is the only fixed effect, called model II, should be preferred. For the sake of comparison, we also fitted the reparametrized model in which R was incorporated as the only fixed effect. The corresponding AIC was 338.620 and though this suggests that the reparametrized model fits the data better than model II, we will keep the latter for the reasons discussed previously.

The estimated standard deviations of the random effects in model II were 0.323 and 0.310 and the estimated correlation coefficient was 0.05. This suggested that the random effects had approximately the same variance and were not correlated. A multiple of the identity matrix was used to model D and the AIC of this reduced model (III) was 338.205, considerably smaller than those of both models I and II. No further reductions were possible, since, if we removed either the lCl random effect
or the IV random effect from model III, we obtained AIC values of 497.968 and 339.799 respectively.

In the next section we explore the use of covariates to explain the cluster-to-cluster variability observed for the ICI and IV random effects.

5. Covariate Modeling

In this section we consider the use of covariates to model random effects variability. This variability can either be related to natural cluster-to-cluster variation, or caused by differences in covariate values between and/or within clusters.

The first questions to be addressed in the covariate modeling process are the determination of which variables are potentially useful in explaining random effects variation and which random effects may have their variability explained by covariates. This is probably best achieved by analyzing plots of the random effects estimates versus the covariates, looking for trends and patterns. The conditional modes of the random effects (Lindstrom and Bates, 1990) were used here for this purpose.

After the candidate covariates have been chosen, a decision has to be made on how to test for their inclusion in the model. The number of extra parameters to be estimated tends to grow considerably with the inclusion of covariates and their associated random effects in the model. If the number of covariates/random effects combinations is large, we suggest using a forward stepwise type of approach in which covariates are included one at a time and the potential importance of the remaining covariates is (graphically) assessed at each step. The decision on whether or not to include a covariate can be based on the AIC of the fits with and without it. Another question that has to be addressed when including a covariate in the model, is which of the new parameters should be random or purely fixed. We suggest using an approach similar to the one described in section 4, for modeling the variance-covariance structure: whenever no prior information is available and convergence is possible, start with a saturated model (in which all new parameters are random) and, by examining the eigenstructure of the estimated $D$ (or $D_{CV}$) matrix, search for plausible structures with fewer parameters. We use the Quinidine data to illustrate this model building approach. We reiterate that any model building strategy is not complete without a careful analysis of residuals and expert advice. In all examples considered here the residual analyses did not indicate departures from the model’s assumptions.

5.1. Quinidine

Figure 2 presents the scatter plots of the conditional modes of the ICI random effect, based on model III of subsection 2.3, versus the available covariates (when the covariate value changed over time, the mode was used). A loess smoother (Cleveland, Grosse and Shyu, 1992) is included in the continuous covariates’ plots to help the visualization of possible trends.

![Figure 2: Conditional modes of the ICI random effect in model III versus available covariates.](image)

Clearance appears to decrease with $\alpha$-1-acid glycoprotein concentration and age and to increase with weight and height. There is also some evidence that clearance decreases with severity of congestive heart failure and is smaller in Blacks than in both Caucasians and Latinas. Clearly the $\alpha$-1-acid glycoprotein concentration is the most important covariate for explaining the ICI cluster-to-cluster variation. A straight line seems adequate to model the observed relationship.

Figure 3 presents the scatter plots of the conditional modes of the IV random effect versus the available covariates. None of the covariates seems helpful in explaining the variability of this random effect and we did not pursue the modeling of its variability any further.

Initially only the $\alpha$-1-acid glycoprotein concentration was included in the model to explain the ICI random effect variation according to a linear model. In the notation of (2.2) this modification of models (3.3) and (3.4) is accomplished by writing

$$ICl_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) \text{glycoprotein}_{ij} \tag{5.1}$$

All parameters were treated as random in this first attempt, called model IV, but the random effects associated with ICI were assumed independent of the IV random effect, preserving the covariance structure of model III. The AIC of this model was 215.796 indicating a considerable gain in goodness-of-fit when compared to model III (AIC of 338.205). Using the same strategy as in section 3 to model the variance-covariance matrix of the random effects, we selected a model in which the ICI random
effect was independent of the IV random effect, the variance-covariance matrix of the $ICl$ random effects was unstructured, but the variances of the intercept term in the $ICl$ random effect, $b_{11}$ in (5.1), and the IV random effect were the same. The AIC of this model (V) was 213.788.

Figures 4 and 5 present the scatter plots of the conditional modes of the $ICl$ random effects in model V versus the available covariates. These plots indicate that the intercept random effect does not vary systematically with any of the covariates, but the slope random effect tends to increase with weight and height and is smaller among Blacks and patients with previous history of congestive heart failure. This suggests an interaction between the effects of these covariates and the $\alpha$-1-acid glycoprotein on the Quinidine clearance. At this point expert advice would be needed to clarify the plausibility of this hypothesis. Since this was not possible here, we proceeded the model building analysis just for the purpose of illustrating the use of the proposed methodology.

Using the forward stepwise approach we included the interactions between $\alpha$-1-acid glycoprotein and weight (as a linear predictor), race (as an indicator variable of black/not black status), and congestive heart failure (as an indicator variable of previous/no previous history of congestive heart failure) in the model, in this order. The same random effect variance-covariance structure as in model V was used in all cases. The corresponding AIC values were respectively 210.117, 204.556, and 199.893. In all three cases substantial reductions in the AIC values were observed. The random effects plots of the last model (with all three interactions with $\alpha$-1-acid glycoprotein included) did not indicate any other candidate covariates to be included in the model and we concluded that the model was adequate.

6. Conclusions

The analysis of the eigenstructure of the estimated variance-covariance matrix ($D$) of the random effects is a useful tool to determine which terms in the model should be random and which should be purely fixed. The estimated $D$ matrix also provides useful information to identify structured variance-covariance patterns. Information criterion statistics, such as the AIC, can be used as guidelines to model selection, but analysis of residuals and consultation with experts in the field of application of the model should also be used.

The goodness-of-fit and interpretability of a mixed effects model can be substantially enhanced through the inclusion of covariates to explain random effects variability. Information criterion statistics can again be used for model selection, together with analysis of residuals and expert advice.

We restricted ourselves here to mixed models in which the cluster errors, $\epsilon$ in model (2.1), were i.i.d., but other covariance structures (e.g. autoregressive processes) can easily be incorporated into mixed models (Chi and Reinsel, 1989; Lindstrom and Bates, 1990). There is usually a trade-off between the number of random effects incorporated in the model and the complexity of the cluster errors covariance structure (Jones, 1990). Further research is needed in that area, especially under a model building...
Figure 5: Conditional modes of $1CI$ slope random effect in model V versus available covariates.

Once a model has been chosen to represent the data, measures of variability for the estimates and confidence regions on the model’s parameters are usually needed for inferential purposes. Maximum likelihood asymptotic theory results can certainly be used at a preliminary stage. These results have not yet been proven for the nonlinear mixed effect model, but have been established for the linear mixed effects model (Pinheiro, 1994). Hence, at least as a first order approximation, asymptotic standard errors and confidence regions based on the normal distribution can be used.

More refined methods, such as likelihood profile traces and contours (Bates and Watts, 1988), can also be used to assess the variability in the estimates, but these will generally be more computationally intensive. A compromise between the asymptotic and profiling methodologies is to use a linear approximation to the loglikelihood, as in Lindstrom and Bates (1990), to calculate the profile traces and contours. This constitutes a considerably less intensive computational problem than profiling the loglikelihood directly.

Another (computationally intensive) alternative to assess the variability in the estimates is to use bootstrap methods (Efron and Tibshirani, 1993) to estimate standard errors and confidence regions.

More research is needed to determine which methods provide the most reliable statistical results and to compare their relative computational performance.

REFERENCES


