Computational Methods for Nonlinear Mixed Models

Douglas Bates

University of Wisconsin - Madison

<Bates@Wisc.edu>

Slides for this presentation are available at
lme4.R-forge.R-project.org/slides/

Joint Statistical Meetings
Washington, DC
August 6, 2009

Outline

Outline

Outline

Outline

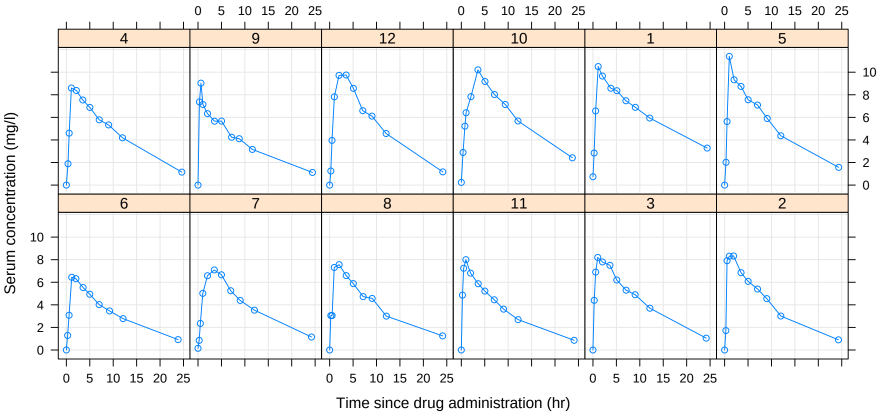
1. Introduction
2. Model definition and an example
3. The penalized least squares problem

Introduction

- Population pharmacokinetics data are often modeled using *nonlinear mixed-effects models* (NLMMs).
- These are *nonlinear* because pharmacokinetic parameters - rate constants, clearance rates, etc. - occur nonlinearly in the model function.
- In statistical terms these are *mixed-effects models* because they involve both *fixed-effects parameters*, applying to the entire population or well-defined subsets of the population, and *random effects* associated with particular experimental or observational units under study.
- Many algorithms for obtaining parameter estimates, usually the *maximum likelihood estimates* (MLEs), for such models have been proposed and implemented.
- Comparing different algorithms is not easy. Even understanding the definition of the model and the proposed algorithm is not easy. We begin by defining the model.

Outline

Theophylline pharmacokinetics



- These are serum concentration profiles for 12 volunteers after injestion of an oral dose of Theophylline, as described in Pinheiro and Bates (2000).

Modeling pharmacokinetic data with a nonlinear model

- These are longitudinal repeated measures data.
- For such data the time pattern of an individual's response is determined by pharmacokinetic parameters (e.g. rate constants) that occur nonlinearly in the expression for the expected response.
- The form of the nonlinear model is determined by the pharmacokinetic theory, not derived from the data.

$$d \cdot k_e \cdot k_a \cdot C \frac{e^{-k_e t} - e^{-k_a t}}{k_a - k_e}$$

- These pharmacokinetic parameters vary over the population. We wish to characterize typical values in the population and the extent of the variation.
- Thus, we associate random effects with the parameters, k_a, k_e and C in the nonlinear model.

Linear and nonlinear mixed-effects models

- For both linear and nonlinear mixed-effects models, we consider the n-dimensional response random variable, $\boldsymbol{\mathcal{Y}}$, whose value, $\boldsymbol{y}$, is observed, and the q-dimensional, unobserved random effects variable, $\boldsymbol{\mathcal{B}}$.
- In the models we will consider $\boldsymbol{\mathcal{B}} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$. The variance-covariance matrix $\boldsymbol{\Sigma}_\theta$ can be huge but it is completely determined by a small number of *variance-component parameters*, $\boldsymbol{\theta}$.
- The conditional distribution of the response, $\boldsymbol{\mathcal{Y}}$, is

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}, \sigma^2 \boldsymbol{I}_n\right)$$

- The conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}$, depends on $\boldsymbol{b}$ and on the fixed-effects parameters, $\boldsymbol{\beta}$, through a *linear predictor* expression, $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}$.
- For a linear mixed model (LMM), $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}$ is exactly the linear predictor. For an NLMM the linear predictor determines the parameter values in the nonlinear model function which then determines the mean.

Transforming to orthogonal random effects

- We never really form $\mathbf{\Sigma}_\theta$; we always work with the *relative covariance factor*, $\mathbf{\Lambda}_\theta$, defined so that

$$\mathbf{\Sigma}_\theta = \sigma^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^\intercal.$$

  Note that we must allow for $\mathbf{\Lambda}_\theta$ to be less that full rank.

- We define a q-dimensional "spherical" or "unit" random-effects vector, $\boldsymbol{\mathcal{U}}$, such that

$$\boldsymbol{\mathcal{U}} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_q\right), \; \boldsymbol{\mathcal{B}} = \mathbf{\Lambda}_\theta \boldsymbol{\mathcal{U}} \Rightarrow \mathsf{Var}(\boldsymbol{\mathcal{B}}) = \sigma^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^\intercal = \mathbf{\Sigma}_\theta.$$

- Setting $\boldsymbol{U}_\theta = \boldsymbol{Z}\mathbf{\Lambda}_\theta$, the linear predictor expression becomes

$$\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Z}\mathbf{\Lambda}_\theta\, \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{U}_\theta\, \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}.$$

The conditional mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$

- Although the probability model is defined from $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}} = \boldsymbol{u})$, we observe $\boldsymbol{y}$, not $\boldsymbol{u}$ (or $\boldsymbol{b}$) so we want to work with the other conditional distribution, $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{y})$.
- The joint distribution of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ is Gaussian with density

$$
\begin{aligned}
f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{y},\boldsymbol{u}) &= f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{y}|\boldsymbol{u})\, f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \\
&= \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\|^2)}{(2\pi\sigma^2)^{n/2}}\;\frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \\
&= \frac{\exp(-\left[\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\|^2+\|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}}
\end{aligned}
$$

- The mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$, of the conditional distribution $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}} = \boldsymbol{y})$ (also the mean in this case of an LMM) is

$$
\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}} \left[\left\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\right\|^2+\|\boldsymbol{u}\|^2\right]
$$

Outline

Minimizing a penalized sum of squared residuals

- An expression like $\left\|\boldsymbol{y} - \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\right\|^2 + \|\boldsymbol{u}\|^2$ is called a *penalized sum of squared residuals* because $\left\|\boldsymbol{y} - \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\right\|^2$ is a sum of squared residuals and $\|\boldsymbol{u}\|^2$ is a penalty on the size of the vector $\boldsymbol{u}$.
- Determining $\tilde{\boldsymbol{u}}_{\theta,\beta}$ as the minimizer of this expression is a *penalized least squares* (PLS) problem. For an LMM it is a *penalized linear least squares problem* that can be solved directly (i.e. without iterating). For an NLMM it is a *penalized nonlinear least squares problem*.
- One way to determine the solution in an LMM is to rephrase it as a linear least squares problem for an extended residual vector

$$\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{U}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2$$

This is sometimes called a *pseudo-data* approach because we create the effect of the penalty term, $\|\boldsymbol{u}\|^2$, by adding "pseudo-observations" to $\boldsymbol{y}$ and to the predictor.

The profiled deviance for LMMs

- We can see that $\tilde{\boldsymbol{u}}_{\theta,\beta}$ satisfies $\left(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{U}_\theta^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ which we solve using the sparse Cholesky decomposition

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T} = \boldsymbol{P}\left(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\boldsymbol{P}^\mathsf{T}$$

  $\boldsymbol{P}$ is a permutation matrix that has practical importance but does not affect the theory. The matrix $\boldsymbol{L}_\theta$ is the sparse, lower-triangular factor.
- Let $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$ be the minimum penalized residual sum of squares, then $\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = \log L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y})$ can be written

$$-2\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = n\log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2} + \log(|\boldsymbol{L}_\theta|^2)$$

- The conditional estimate of σ^2 is

$$\widehat{\sigma^2}(\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}$$

  producing the *profiled deviance*

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$$

Profiling the deviance with respect to β for LMMs

- In a LMM the deviance depends on β only through $r^2(\boldsymbol{\theta}, \boldsymbol{\beta})$ we can obtain the conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, by extending the PLS problem to

$$r^2(\boldsymbol{\theta}) = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left[\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}_\theta\, \boldsymbol{u} \right\|^2 + \left\| \boldsymbol{u} \right\|^2 \right]$$

with the solution satisfying the equations

$$\begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{U}_\theta + \boldsymbol{I}_q & \boldsymbol{U}_\theta^\intercal \boldsymbol{X} \\ \boldsymbol{X}^\intercal \boldsymbol{U}_\theta & \boldsymbol{X}^\intercal \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_\theta \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_\theta^\intercal \boldsymbol{y} \\ \boldsymbol{X}^\intercal \boldsymbol{y}. \end{bmatrix}$$

- The profiled deviance, which is a function of $\boldsymbol{\theta}$ only, is

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}_\theta|^2) + n \left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta})}{n} \right) \right]$$