Mixed-Effects Non-Linear and Generalized Linear Models

Douglas Bates

University of Wisconsin - Madison
<Bates@Wisc.edu>

DAEW02 Workshop, Issac Newton Institute
August 9, 2011

Outline

Outline

Outline

Outline

Outline

Outline

Outline

Mixed-effects Models

- From the statistical point of view, mixed-effects models involve two types of coefficients or "effects":

  Fixed-effects parameters, which are characteristics of the entire population or well-defined subsets of the population

  Random effects, which are characteristics of individual experimental or observational units.

- In the probability model we consider the distribution of two vector-valued random variables: $\mathcal{Y}$, the n-dimension response vector and $\mathcal{B}$, the q-dimensional vector of random effects.
- The value, $\boldsymbol{y}_{\text{obs}}$, of $\mathcal{Y}$ is observed; the value of $\mathcal{B}$ is not.

Distributions of the random variables

- In the probability model we specify the unconditional distribution of $\mathcal{B}$ and the conditional distribution of $\mathcal{Y}$, given $\mathcal{B} = \boldsymbol{b}$.
- Because the random effects, $\mathcal{B}$, are unobserved, the assumed distribution is kept simple. For most of the models that we will describe we assume

$$\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

  where $\boldsymbol{\Sigma}$ is a parameterized, positive semi-definite symmetric matrix.
- In the conditional distribution, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$, the value $\boldsymbol{b}$ changes only the conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}}$, and does so through a *linear predictor* expression

$$\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$$

  where $\boldsymbol{\beta}$ is a p-dimensional fixed-effects vector and the model matrices, $\boldsymbol{X}$ and $\boldsymbol{Z}$, are of the appropriate dimension.

Linear Mixed Models

- In a linear mixed model (LMM) the distributions of $\mathcal{Y}$ and $\mathcal{B}$ are both Gaussian and the conditional mean is the linear predictor,

$$\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}.$$

- More explicitly

$$(\mathcal{Y}|\mathcal{B} = \boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma^2\boldsymbol{I}_n)$$

and

$$\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^\mathsf{T})$$

In the expression $\sigma^2\boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^\mathsf{T}$ the scale parameter, σ, is the same as that in the expression for $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$, and $\boldsymbol{\Lambda}_\theta$ is the parameterized *relative covariance factor*.

Generalized linear mixed models

- In a generalized linear mixed model (GLMM) the conditional distribution, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$ can be other than Gaussian. Common choices are Bernoulli for binary response data and Poisson for count data. Some of the theory works best when this distribution is from the exponential family.
- Because each element of $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}}$ may restricted to an interval, (e.g. $(0, 1)$ for the Bernoulli or $(0, \infty)$ for the Poisson), the conditional mean is expressed as a non-linear function, $\boldsymbol{g}^{-1}$, called the *inverse link*, of the linear predictor, $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zb}$

$$\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}) = \boldsymbol{g}^{-1}(\boldsymbol{X\beta} + \boldsymbol{Zb})$$

- The inverse link is defined by applying a scalar inverse link function, g^{-1}, componentwise, $\mu_i = g^{-1}(\eta_i)$. Thus the Jacobian matrix, $d\boldsymbol{\mu}/d\boldsymbol{\eta}$, is diagonal.

Generalized linear mixed models (cont'd)

- We must be more explicit about the multivariate distribution, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$.
- Components of $\mathcal{Y}$ are *conditionally independent*, given $\mathcal{B} = \boldsymbol{b}$.
- In many common cases this means that the conditional mean entirely determines the conditional distribution.
- It is a common misconception that the variance-covariance of $\mathcal{Y}$ can be modelled separately from the mean. With a Gaussian conditional distribution you can separately model the mean and the variance. With most other conditional distributions you can't.
- Another common misconception is that there is an advantage in writing the conditional distribution in a "signal"+ "noise" form like

$$(\mathcal{Y}|\mathcal{B} = \boldsymbol{b}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

for the Gaussian case. This doesn't gain you anything and induces considerable confusion.

Nonlinear mixed models

- The nomenclature here is a bit tricky. Even though a GLMM can, and often does, involve a nonlinear inverse link function, g^{-1}, we reserve the term *nonlinear mixed-effects model* (NLMM) for cases where the transformation from linear predictor to conditional mean involves a nonlinear model function separate from the inverse link.
- The nonlinear model function, $h(\boldsymbol{x}_i, \boldsymbol{\phi}_i)$, is usually a *mechanistic model* (i.e. based on an external theory of the mechanism under study) as opposed to an *empirical model* derived from the data.
- For example, in pharmacokinetics, a two-compartment open model for the serum concentrations of a drug administered orally at $t = 0$ is

$$h(\boldsymbol{x}_i, \boldsymbol{\phi}_i) = k_e \cdot k_a \cdot C \frac{e^{-k_e t_i} - e^{-k_a t_i}}{k_a - k_e}$$

where k_a is the absorption rate constant, k_e is the elimination rate constant and C is the clearance; the covariate vector $\boldsymbol{x}_i$ for the ith observation is t_i and the nonlinear parameter vector $\boldsymbol{\phi}_i$ is (k_a, k_e, C).

Nonlinear mixed models (cont'd)

- In the basic nonlinear mixed model, the conditional distribution, $\mathcal{Y}|\mathcal{B} = \boldsymbol{b}$, is a spherical Gaussian

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

- A further extension, of course, is to allow for a generalized nonlinear mixed model (GNLMM) in which the conditional mean is a nonlinear function (in addition to an inverse link) of the linear predictor and the conditional distribution is non-Gaussian.
- There are important applications for such models in what is called *item-response theory* that provides models for correct/incorrect answers on objective exams according to characteristics of the items (difficulty, discrimination, threshold probability for guessing) and characteristics of the subjects (ability).

Linear and nonlinear mixed-effects models

- The two "non-generalized" model forms are sufficiently alike that it is worthwhile considering them together. Both can be written as

$$(\mathcal{Y}|\mathcal{B}=\boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}}, \sigma^2\boldsymbol{I}_n), \quad \mathcal{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathcal{N}(\sigma^2\mathbf{0}, \boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^\mathsf{T})$$

  It is only the relationship between the linear predictor, $\boldsymbol{\eta}$, and the conditional mean, $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{B}}$, that differs.
- The joint density for $\mathcal{Y}$ and $\mathcal{B}$ is

$$f_{\mathcal{Y},\mathcal{B}}(\boldsymbol{y},\boldsymbol{b}) = f_{\mathcal{Y}|\mathcal{B}}(\boldsymbol{y}|\boldsymbol{b})\, f_{\mathcal{B}}(\boldsymbol{b})$$

  providing the marginal density

$$f_{\mathcal{Y}}(\boldsymbol{y}) = \int_{\mathbb{R}^q} f_{\mathcal{Y},\mathcal{B}}(\boldsymbol{y},\boldsymbol{b})\, d\boldsymbol{b}$$

  and the likelihood

$$L(\boldsymbol{\beta},\boldsymbol{\theta},\sigma|\boldsymbol{y}) = f_{\mathcal{Y}}(\boldsymbol{y}_{\text{obs}}).$$

"Spherical" random effects

- At this point we introduce a linear transformation, determined by $\boldsymbol{\Lambda}_\theta$, of the random effects. Recall that $\boldsymbol{\Lambda}_\theta$ can be singular (it is only required to be positive semi-definite). The maximum likelihood estimates (mle's) of variance components can be zero.
- Even if the estimates are not on the boundary of the parameter space, we may need to evaluate on the boundary while optimizating.
- This is why algorithms based on estimating the precision matrix, $\boldsymbol{\Sigma}^{-1}$, (e.g. EM algorithms) or requiring its value (Henderson's mixed model equations) run into problems.
- You can evaluate the likelihood on the boundary – you just need to be careful how you evaluate it.
- We define a "spherical" random effects vector, $\mathcal{U}$,

$$\mathcal{B} = \boldsymbol{\Lambda}_\theta \mathcal{U}, \quad \mathcal{U} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_q)$$

with linear predictor, $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u}$.

Joint densities and conditional modes

- The joint density function for $\mathcal{Y}$ and $\mathcal{U}$, which is the quantity in the integrand for the likelihood, is

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

- This expression, evaluated at $\boldsymbol{y}_{\text{obs}}$ is the unnormalized conditional density of $\mathcal{U}$ given $\mathcal{Y}=\boldsymbol{y}_{\text{obs}}$. (In fact, the inverse of the normalizing factor is exactly the likelihood.)
- The *conditional mode*, $\tilde{\boldsymbol{u}}(\boldsymbol{y}_{\text{obs}})$, of the random effects is the solution of the penalized least squares (PLS) problem

$$\tilde{\boldsymbol{u}}(\boldsymbol{y}_{\text{obs}}) = \arg\min_{\boldsymbol{u}}\left(\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\|^2+\|\boldsymbol{u}\|^2\right)$$

Solving the linear PLS problem

- For a linear mixed model the PLS problem is a penalized linear least squares problem and the conditional mode is also the conditional mean of $\mathcal{U}|\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$. For a nonlinear model the PLS problem is a penalized nonlinear least squares problem.
- In the linear case there is a direct solution to the PLS problem. In fact, we can simultaneously determine $\tilde{\boldsymbol{u}}$ and $\widehat{\boldsymbol{\beta}}_\theta$, the conditional estimate of $\boldsymbol{\beta}$, as the minimizers of

$$r_\theta^2 = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \right]$$

which are the solutions to the system

$$\begin{bmatrix} \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q & \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}} \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \end{bmatrix}.$$

Use of the sparse Cholesky factor

- Taking into account that the dimensions of $\boldsymbol{Z}$ can be very large indeed, the equations for the PLS solutions would be interesting but not terribly useful, except that $\boldsymbol{Z}$ (and $\boldsymbol{\Lambda}_\theta$) are also very sparse.
- The system matrix, especially the part $\boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q$ is positive definite, even when $\boldsymbol{\Lambda}_\theta$ is singular.
- Determining the sparse Cholesky factor, $\boldsymbol{L}_\theta$, which is a sparse lower triangular matrix such that

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T} = \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q$$

  is a well-understood process for which high quality, effective software is available.
- Like most operations on sparse matrices, the sparse Cholesky factorization is performed in two phases: a *symbolic phase* in which the positions of the non-zeros in the result are determined, and a *numeric phase* in which the actual numeric values are calculated. The symbolic phase need only be done once.

The profiled deviance and REML criterion

- Given a value of $\boldsymbol{\theta}$ we determine the sparse Cholesky factor, $\boldsymbol{L}_\theta$, the conditional mode, $\tilde{u}_\theta$, of the random effects and the conditional estimates, $\widehat{\boldsymbol{\beta}}_\theta$ and $\widehat{\sigma^2}_\theta$ of the other parameters, providing the *profiled deviance* as a function of $\boldsymbol{\theta}$ only.

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

- The REML criterion is

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta}$$

and the profiled REML criterion can be evaluated as

$$-2\tilde{\ell}_R(\boldsymbol{\theta}) = \log(|\boldsymbol{L}|^2) + \log(|\boldsymbol{R}_x|^2) + (n-p)\left[1 + \log\left(\frac{2\pi r_\theta^2}{n-p}\right)\right]$$

where $\boldsymbol{R}_{\boldsymbol{X}}$ is the $p \times p$ (usually dense) Cholesky factor in the full decomposition of the system matrix for the PLS problem.

Laplace approximation to the deviance for an NLMM

- For an NLMM, the PLS problem becomes penalized nonlinear least squares, which usually requires an iterative solution, such as using the Gauss-Newton algorithm.
- We can determine the solution with respect to $\boldsymbol{u}$ only or simultaneously with respect to $\boldsymbol{u}$ and $\boldsymbol{\beta}$. In the latter case, the $\boldsymbol{\beta}$ optimizer is close to but not necessarily the same as the conditional estimate $\widehat{\boldsymbol{\beta}}_\theta$.
- The *Laplace approximation* to the profiled deviance is

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

where r_θ^2 is the minimum penalized residual sum of squares and $\boldsymbol{L}_\theta$ is the sparse Cholesky factor at the PNLS solition. If $\boldsymbol{\beta}$ is not optimized during the PNLS problem then these quantities should be indexed by $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Adaptive Gauss-Hermite quadrature

- The Laplace approximation involves approximating the unnormalized density of $\mathcal{U}|\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$ by a multivariate Gaussian that matches the mode and the second moment at the mode.
- Gauss-Hermite quadrature provides weights and abscissa values to evaluate scalar integrals of the form $\int_{\mathbb{R}} f(x)e^{-x^2}\,dx$ as a linear combination of function values. Extensions to multivariate integrals, evaluating either on grids or on spherical patterns exist but are only suitable for low dimensions.
- If the integral of the unnormalized conditional density can be factored into the product of low-dimensional integrals then these can be evaluated more accurately using Gauss-Hermite quadrature.
- This process is called *adaptive Gauss-Hermite quadrature* (AGQ) because the quadrature points are evaluated taking into account the conditional mode and the second moment of the unnormalized density at the conditional mode.

When can AGQ be used?

- The random effects are associated with the levels of one or more factors, called the *grouping factors*, in the data. In the simple case where there is only one grouping factor (e.g. random effects for `Subject` only) the observations can be grouped according to the levels of this single grouping factor.
- Conditional independence in the distribution $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}$ and independence of components in $\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_q)$ allows the multivariate integral to be expressed as the product of scalar or low-dimensional integrals.

Maximum likelihood estimates for GLMMs

- GLMMs also can have a nonlinearity in the transformation from $\boldsymbol{\eta}$, the linear predictor, to $\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}$, induced by the inverse link function.
- Furthermore, in a GLMM changing the conditional mean can change the conditional variance of $\mathcal{Y}$ given $\mathcal{U} = \boldsymbol{u}$ and we account for this by using weighted least squares.
- Some complications of notation can arise because $\mathcal{Y}|\mathcal{U} = \boldsymbol{u}$ is often a discrete distribution. Nonetheless, $\mathcal{U}$ is always continuous and the unscaled conditional density of $\mathcal{U}|\mathcal{Y} = \boldsymbol{y}_{\text{obs}}$ is well-defined.
- The iteratively reweighted least squares (IRLS) algorithm for determining the mle's in a generalized linear model (GLM) is modified to PIRLS for determining the conditional mode, $\tilde{\boldsymbol{u}}$, in a GLMM. The Laplace and AGQ approximations follow as for NLMMs.

Taxonomy of mixed-model forms

- In a linear mixed model the distribution of the response, given the random effects, is a multivariate Gaussian whose mean is mean is the linear predictor, $\boldsymbol{X\beta} + \boldsymbol{Zb}$.
- In a generalized linear mixed model, the conditional distribution is non-Gaussian with a mean that can be a transformation of the linear predictor. (For historical reasons this function is called the "inverse link".) The Rausch IRT model is an example.
- In a nonlinear mixed model the conditional distribution is Gaussian but the mean function is nonlinear in one or more of the fixed-effects parameters or the random effects (or both).
- In a generalized nonlinear mixed model the conditional distribution is non-Gaussian and the mean function is nonlinear in parameters or random effects (beyond the nonlinearity of the inverse link).
- The inner optimization problem for each of these cases is PLS (penalized linear least squares), PIRLS (penalized iteratively reweighted least squares), PNLS (penalized nonlinear least squares) and PIRNLS.