Statistical Methods for Correlated Data

Likelihood Inference for Generalized Linear Mixed Models

Michele Guindani

Department of Biostatistics
UCLA

- In GLMMs, there are three distinct sets of parameters for which inference may be required: fixed effects $\boldsymbol{\beta}$, variance components $\boldsymbol{\alpha}$, and random effects $\boldsymbol{b} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^{\mathrm{T}}$.
- Objective: maximize the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denote the variance components in $\boldsymbol{D}$ and the scale parameter ϕ (if present).
- The likelihood is obtained by integrating $[\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$ from the model:

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{m} \int p\left(\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{b}_i\right) \times p\left(\boldsymbol{b}_i | \boldsymbol{\alpha}\right) d\boldsymbol{b}_i$$

- For non-Gaussian GLMMs, these integrals are not available in closed form.
- The resulting (approximated) likelihood may not be regular, so the second derivatives may be difficult to determine, so the Newton-Raphson method cannot be directly used ⇝ quasi-Newton approach (will discuss more later methods of estimating parameters)

Procedures of maximization and hypothesis testing on fixed effects

- Given various estimating procedures for GLMMs, β is asymptotically normal, given the large-sample approximation to the joint distribution of parameters, with mean $\mathbf{0}$ and variance $\boldsymbol{I}_{\beta\beta}^{-1}$.
- Briefly,

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} \rightsquigarrow \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{\beta\beta}\boldsymbol{I}_{\beta\alpha} \\ \boldsymbol{I}_{\alpha\beta}\boldsymbol{I}_{\alpha\alpha} \end{bmatrix}^{-1} \right)$$

- In general, $\boldsymbol{I}_{\beta\alpha} \neq \mathbf{0}$ (compare with equation 8.20 in the textbook for LME) and so with GLMMs we cannot have REML estimates
☞ consistency requires the specification of both mean and variance parameters

Hypothesis testing

- If the asymptotic distribution is valid, we can use standard approaches for conducting hypothesis testing on the fixed effects.
- Example, Wald test: for testing $\mathrm{H}_0 : \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ we can use

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p$$

  where $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$ is the asymptotically consistent estimator for the expected information matrix $\boldsymbol{I}(\boldsymbol{\beta})$

Hypothesis testing

- If the asymptotic distribution is valid, we can use standard approaches for conducting hypothesis testing on the fixed effects.
- Example, Wald test: for testing $\mathrm{H}_0 : \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ we can use

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p$$

  where $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$ is the asymptotically consistent estimator for the expected information matrix $\boldsymbol{I}(\boldsymbol{\beta})$
- Example, LRT: can be used to test a null reduced model equation, as the likelihood ratio test statistics is asymptotically distributed as a χ^2 with degrees of freedom = to the difference in number of parameters in the model
- Between the likelihood ratio and the Wald statistics, the likelihood ratio statistic is considered to provide more reliable test results although it is computationally more complex. Therefore, when possible the likelihood ratio test is preferable to test a null hypothesis on the fixed effects.

Hypothesis testing on the variance components

- When the significance test is performed on the variance-covariance components of the random effects, and the test is at the boundary of the parameter space, e.g. $\mathrm{H}_0 : \sigma^2_{b_{\tilde{i}}} = 0$ and $\mathrm{H}_1 : \sigma^2_{b_{\tilde{i}}} > 0$ where σ_{b_i} is the i-th diagonal element in $\boldsymbol{D}$, then the conventional tests, such as the likelihood ratio and the Wald statistics, cannot be directly applied to test these hypotheses.
- Since the $\boldsymbol{D}$ matrix in GLMMs is generally specified to follow normality, one can still use the procedures described for LMEs and use mixtures of χ^2 distributions to test the null

Methods for estimating parameters in GLMMs

- In the analysis of nonnormal longitudinal data, such as proportions or counts, numerical integration or Bayes-type techniques are required to conduct a full ML analysis.
- There are a variety of methods used to approximate the relevant likelihood quantities:
  1. Gaussian quadrature rules
  2. Laplace approximation
  3. Penalized quasi-likelihood (PQL)
  4. Marginal quasi-likelihood (MQL)
  5. Monte Carlo EM
  6. Markov Chain Monte Carlo (MCMC)

Gaussian quadrature methods

- A widely used approximation method of the integrals in GLMMs is Gaussian quadrature
- Suppose that there is a known, smooth function $f(z)$ and a probability density function $p(z)$. The function $f(z)$ can then be integrated against $p(z)$. The corresponding quadrature rule is

$$\int_{-\infty}^{\infty} f(z)p(z)\mathrm{d}z \approx \sum_{q=1}^{Q} w_q f\left(z_q\right)$$

where Q represents the number of quadrature points, w_q is the quadrature weight ($q = 1, \ldots, Q$), and z_q is the abscissas, statistically referred to as a node.

Gaussian quadrature methods

- In GLMMs, $p(z)$ is represented by the distribution of the random effects. However, in Gaussian quadratures, $p(z)$ is standard normal, so the the random effects may first be standardized

$$\tilde{\boldsymbol{b}}_i = \boldsymbol{D}^{-1/2}\boldsymbol{b}_i$$

  so that $\tilde{\boldsymbol{b}} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. Correspondingly, one needs to consider $\boldsymbol{\eta}_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{D}^{1/2}\boldsymbol{b}_i$

- Given the above specification, the likelihood contribution for subject i is

$$\begin{aligned} f_i\left(\boldsymbol{y}_i|\boldsymbol{\beta}, \boldsymbol{D}, \phi\right) &= \int \prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi\right) f\left(\boldsymbol{b}_i|\boldsymbol{G}\right) \mathrm{d}\boldsymbol{b}_i \\ &= \int \prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\boldsymbol{\beta}, \hat{\boldsymbol{b}}_i, \boldsymbol{D}, \phi\right) f\left(\hat{\boldsymbol{b}}_i\right) \mathrm{d}\hat{b}_i \end{aligned}$$

- A widely used algorithm is the Gauss-Hermite quadrature (details omitted)

Laplace Approximation

- The Laplace method is a statistical technique for approximating integrals of the form

$$\int_{\hat{a}}^{\hat{b}} \exp[Nf(z)]\mathrm{d}z$$

where $f(z)$ is some known, smooth function that is unimodal with a maximum at z_0 and N is the sample size.

- Therefore, z_0 is assumed to be the only point satisfying

$$\Delta f(z) = \left[\frac{\partial f(z)}{\partial z_{\hat{a}}}, \frac{\partial f(z)}{\partial z_{\hat{a}+1}}, \ldots, \frac{\partial f(z)}{\partial z_{\hat{b}}}\right] = 0$$

- It is also assumed that the Hessian matrix of $f(z)$ at z_0

$$\boldsymbol{H}_0' = \left[\frac{\partial^2 f(z)}{\partial z_i \partial z_j}\right]\Bigg|_{z=z_0}$$

is positive definite

Laplace Approximation

- According to Taylor's theorem, $f(z)$ can be further expanded to the expression

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \frac{1}{2} f''(z_0)(z - z_0)^2 + O\left[(z - z_0)^3\right]$$

where $O\left[(z - z_0)^3\right] \xrightarrow{p} 0$.

- As $f'(z_0) = 0$ by assumption, $f(z)$ can be approximated to the quadratic form

$$f(z) = f(z_0) + \frac{1}{2} f''(z_0)(z - z_0)^2$$

- Then, assuming $f''(z_0) < 0$,

$$\int_{\hat{a}}^{\hat{b}} \exp[Nf(z)]\mathrm{d}z \approx \exp\{N f(z_0)\} \int_{\hat{a}}^{\hat{b}} \exp\left\{-\frac{N}{2}\{-f''(z_0)\}(z - z_0)^2\right\} dz$$

which is a Normal kernel $(N(z_0, \{-f''(z_0)\}^{-1}/N))$.

Laplace approximation

- Hence,

$$\int_{\hat{a}}^{\hat{b}} \exp[Nf(z)]\mathrm{d}z \approx$$

$$\exp\{N\, f(z_0)\} \sqrt{\frac{2\pi N}{\{-f''\,(z_0)\}}} \int_{\hat{a}}^{\hat{b}} \phi(z|z_0, \{-f''\,(z_0)\}^{-1}/N)\, dz$$

Laplace approximation

- Hence,

$$\int_{\hat{a}}^{\hat{b}} \exp[Nf(z)]\mathrm{d}z \approx$$

$$\exp\{N\, f(z_0)\} \sqrt{\frac{2\pi N}{\{-f''\,(z_0)\}}} \int_{\hat{a}}^{\hat{b}} \phi(z|z_0, \{-f''\,(z_0)\}^{-1}/N)\, dz$$

- The Laplace approx can be easily extended to the longitudinal setting.

$$\begin{aligned} p(\boldsymbol{y}) &= \prod_{i=1}^{m} p\,(\boldsymbol{y}_i) = \prod_{i=1}^{m} \int p\,(\boldsymbol{y}_i|\boldsymbol{b}_i)\, p\,(\boldsymbol{b}_i)\, \mathrm{d}\boldsymbol{b}_i \\ &= \prod_{i=1}^{m} \int \exp\left\{\log\left[p\,(\boldsymbol{y}_i|\boldsymbol{b}_i)\, p\,(\boldsymbol{b}_i)\right]\right\} \mathrm{d}\boldsymbol{b}_i \\ &= \prod_{i=1}^{N} \int \exp\left\{\sum_{j=1}^{n_i} \log\left[p\,(y_{ij}|\boldsymbol{b}_i)\right] + n_i \log\left[p\,(\boldsymbol{b}_i)\right]\right\} \mathrm{d}\boldsymbol{b}_i \end{aligned}$$

Laplace Approximation

- After some algebra, the Laplace approximation to the marginal log-likelihood can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \hat{\boldsymbol{b}}, \boldsymbol{y}) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \log \left[p\left(y_{ij} | \hat{\boldsymbol{b}}_i \right) \right] + n_i \log \left[p\left(\hat{\boldsymbol{b}}_i \right) \right] \right.$$
$$\left. + \frac{1}{2} n_{b_i} \log(2\pi) - \log \left| -\frac{1}{2} n_i f''\left(\beta, \boldsymbol{\alpha}; \hat{\boldsymbol{b}}_i \right) \right| \right]$$

where n_{b_i} is the dimension of the random effect $\boldsymbol{b}_i$

Laplace Approximation

- After some algebra, the Laplace approximation to the marginal log-likelihood can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}; \hat{\boldsymbol{b}}, \boldsymbol{y}) = \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \log \left[p\left(y_{ij} | \hat{\boldsymbol{b}}_i \right) \right] + n_i \log \left[p\left(\hat{\boldsymbol{b}}_i \right) \right] \right.$$
$$\left. + \frac{1}{2} n_{b_i} \log(2\pi) - \log \left| -\frac{1}{2} n_i f''\left(\beta, \boldsymbol{\alpha}; \hat{\boldsymbol{b}}_i \right) \right| \right]$$

  where n_{b_i} is the dimension of the random effect $\boldsymbol{b}_i$

- Maximizing the above log-likelihood function yields the MLE of the Laplace method parameters.
- This is the default in glmer. The function finds the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, using a conditional maximum of $\boldsymbol{b}$ (obtained as explained in the following slides).

Penalized Iterated Reweighted Least Squares (PIRLS)

- For the Laplace approximation, we need to compute the conditional modes of the random effects:

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \text{argmax}_b \left\{ p(y|\boldsymbol{\beta}, \boldsymbol{b})\, f(b, \boldsymbol{D}(\boldsymbol{\alpha})) \right\}$$

  Conditional modes are calculated using Penalized Iterated Reweighted Least Squares (PIRLS).

- This will be very similar to the IRLS procedure used in the GLM case except that we will add a penalty term:

$$\tilde{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \text{argmax}_b \left\{ \log p(y|\beta, b) - \frac{b^T \boldsymbol{D}^{-1}(\boldsymbol{\alpha}) b}{2} \right\}$$

- To de-clutter notation write b and $\boldsymbol{\beta}$ in a single vector $\boldsymbol{\mathcal{B}}^\top = \left(\mathbf{b}^\top, \boldsymbol{\beta}^\top\right)$, and define corresponding model matrix and precision matrix

$$\mathcal{X} = (\mathbf{Z}, \mathbf{X}) \text{ and } \mathbf{S} = \begin{bmatrix} \boldsymbol{D}(\boldsymbol{\alpha})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

- Then, one can show that the Hessian of the penalized likelihood can be written as $-\mathcal{X}^\top \mathbf{W}\mathcal{X}/\phi - \mathbf{S}$, where $\mathbf{W}$ is the diagonal weight matrix (Fisher or full Newton) implied by $\hat{\boldsymbol{\mu}}$ (e.g. Fisher weights in GLM: $\mathbf{W} = \text{diag}\,(w_i)$ where $w_i = 1/\left\{g'\left(\mu_i\right)^2 V\left(\mu_i\right)\right\}$)
- Let $\mathbf{G} = \text{diag}\left\{g'\left(\mu_i\right)/\alpha\left(\mu_i\right)\right\}$. Then, a single Newton update step has the form

$$\mathcal{B}^{[k+1]} = \mathcal{B}^{[k]} + \left(\mathcal{X}^\top \mathbf{W}\mathcal{X} + \phi\mathbf{S}\right)^{-1}\left\{\mathcal{X}^\top \mathbf{W}\mathbf{G}(\mathbf{y} - \hat{\boldsymbol{\mu}}) - \phi\mathbf{S}\mathcal{B}^{[k]}\right\}$$

- Substituting $\mathcal{B}^{[k]} = \left(\mathcal{X}^\top \mathbf{W}\mathcal{X} + \phi\mathbf{S}\right)^{-1}\left(\mathcal{X}^\top \mathbf{W}\mathcal{X} + \phi\mathbf{S}\right)\mathcal{B}^{[k]}$ and re-arranding, then we obtain

$$\mathcal{B}^{[k+1]} = \left(\mathcal{X}^\top \mathbf{W}\mathcal{X} + \phi\mathbf{S}\right)^{-1}\mathcal{X}^\top \mathbf{W}\left\{\mathbf{G}(\mathbf{y} - \hat{\boldsymbol{\mu}}) + \mathcal{X}\mathcal{B}^{[k]}\right\}$$

Penalized Iterated Reweighted Least Squares (PIRLS)

- The last expression can be seen as the minimizer of the penalized weighted least squares objective

$$\|\mathbf{z} - \mathcal{X}\mathcal{B}\|_W^2 + \phi\mathcal{B}^\top \mathbf{S}\mathbf{B} = \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|_W^2 + \phi\mathbf{b}^\top \boldsymbol{\psi}_\theta^{-1}\mathbf{b}$$

where $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$ with $\hat{\eta}_i = g(\hat{\mu}_i)$

Penalized Iterated Reweighted Least Squares (PIRLS)

- The last expression can be seen as the minimizer of the penalized weighted least squares objective

$$\|\mathbf{z} - \mathcal{X}\mathcal{B}\|_W^2 + \phi\mathcal{B}^\top \mathbf{S}\mathbf{B} = \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|_W^2 + \phi\mathbf{b}^\top \boldsymbol{\psi}_\theta^{-1}\mathbf{b}$$

  where $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i$ with $\hat{\eta}_i = g(\hat{\mu}_i)$

- PIRLS algorithm:

  1. Initialize $\hat{\mu}_i = \hat{\eta}_i$
  2. Compute pseudo-data $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$, and iterative weights $w_i = \alpha(\hat{\mu}_i) / \left\{ g'(\hat{\mu}_i)^2 V(\hat{\mu}_i) \right\}$ where $\alpha(\mu_i)$ is either 1 (Fisher scoring) or
     $= 1 + (y_i - \mu_i)\{V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i)\}$ (Full Newton-Rahpson)
  3. Find $\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}$, to minimize the weighted least squares objective

$$\|\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|_W^2 + \phi\mathbf{b}^\top \boldsymbol{D}(\alpha)^{-1}\mathbf{b}$$

  4. update $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ until convergence

Penalized quasi-likelihood methods

- Using Laplace + PIRLS one obtains estimates of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{b}$. However, the nested optimization required to optimize the Laplace approximate profile is somewhat involved and can be computationally costly.
- As an alternative one can use the PQL to estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{b}$ and $\boldsymbol{\alpha}$ at once using working mixed model. The PQL method estimates the fixed and random effects parameters separately, and then iterate until a convergence criterion is met.

Penalized quasi-likelihood methods

- First note that, given a current estimate $\tilde{\boldsymbol{b}}$ of $\boldsymbol{b}$ (e.g., empirical Bayes estimate), one can consider

$$\mathrm{E}(\boldsymbol{y}|\boldsymbol{b}) = \boldsymbol{\mu} = g^{-1}\left(\boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}\right) = g^{-1}(\boldsymbol{\eta})$$

- Let $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{b}}$ be the current estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}$. According to Wolfinger and O'Connell (1993), the first-order Taylor series expansion of μ about $\tilde{\beta}$ and $\tilde{\boldsymbol{b}}$ yields approximation of mean

$$\mathrm{g}^{-1}(\eta) = \mathrm{g}^{-1}(\tilde{\eta}) + \tilde{\Delta}\boldsymbol{X}'(\beta - \tilde{\boldsymbol{\beta}}) + \tilde{\Delta}\boldsymbol{Z}'(\boldsymbol{b} - \tilde{\boldsymbol{b}}) + \boldsymbol{error}$$

where $\tilde{\Delta}$ is the diagonal matrix of the derivatives of the conditional mean evaluated at the expansion, mathematically defined by $\tilde{\Delta} = \left[\frac{\partial g^{-1}(\eta)}{\partial \eta}\right]_{\tilde{\beta},\tilde{b}}$

Penalized quasi-likelihood methods

- By rearranging terms, the following expression is derived:

$$\tilde{\Delta}^{-1}\left[\mu - \mathrm{g}^{-1}(\tilde{\eta})\right] + X'\tilde{\beta} + Z\tilde{b} = X'\beta + Zb$$

- We can see the left-hand side as the expected value of the pseudo-response variable conditionally on b

$$\tilde{\Delta}^{-1}\left[\boldsymbol{y} - \mathrm{g}^{-1}(\tilde{\boldsymbol{\eta}})\right] + \boldsymbol{X}'\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}'\tilde{\boldsymbol{b}} \equiv \tilde{\boldsymbol{y}}$$

- $\tilde{y}$ is assumed to be normally distributed, i.e.

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{Z}'\boldsymbol{b} + \tilde{\boldsymbol{\varepsilon}}$$

with pseudo-error term $\tilde{\varepsilon} = \tilde{\Delta}^{-1}\left[\tilde{y} - \mathrm{g}^{-1}(\tilde{\eta})\right]$ mean-zero normal with variance $\mathrm{var}(\tilde{\varepsilon}|\boldsymbol{\beta}, \boldsymbol{b}) = \phi\,\tilde{\Delta}^{-1}V_{y_i|\boldsymbol{\beta},\boldsymbol{b}_i}(\boldsymbol{\mu}_i)\tilde{\Delta}^{-1}$

Penalized quasi-likelihood methods

- Then, the estimation of parameters can be performed by following the standard procedures for fitting a linear mixed model: given starting values of $\boldsymbol{\beta}, \boldsymbol{D}$ and ϕ in the marginal likelihood, the empirical Bayes estimates can be computed for $\boldsymbol{b}$, thereby generating the pseudo-data $\tilde{\boldsymbol{y}}$. Consequently, the approximate linear mixed model can be fitted, yielding updated estimates for $\boldsymbol{\beta}, \boldsymbol{D}$, and ϕ. The iterative process continues until a convergence criterion is reached.
- In many cases PQL works surprisingly well,is highly effective and produces estimates very close to the full Laplace approximation based estimates. However, for some types of data it is problematic.
  - for modelling over-dispersed low-mean count data
  - for binary data (in standard implementations of the LMMs, the algorithm estimates ϕ even if it is known; for binary data, there is no real information to estimate ϕ)
  - if there are only a few observations per random effect, the central limit theorem justification for the method is not valid
  - A final general disadvantage is that we do not have access to the profile likelihood for the model itself, so that LRT and AIC based on those is not possible.