Statistical Methods for Correlated Data

Likelihood Inference for Linear Mixed Models

Michele Guindani

Department of Biostatistics
UCLA

Inference for LMMs

- We consider methods for inference in LMMs:

  $$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

  Standard methods of estimation in linear (better, gaussian) mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML) methods

Inference for LMMs

- We consider methods for inference in LMMs:

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

  Standard methods of estimation in linear (better, gaussian) mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML) methods

- In order to conduct likelihood inference, we need distributional assumptions on the random and measurement error terms. As discussed previously, we assume

$$\epsilon_i | \sigma^2_\epsilon \sim_{iid} \mathbf{N}_{n_i}\left(\mathbf{0}, E_i\right) \quad \text{and} \quad \boldsymbol{b}_i | \boldsymbol{D} \sim_{iid} \mathbf{N}_{q+1}(\mathbf{0}, \boldsymbol{D})$$

  and we further assume $E_i = \sigma^2_\epsilon \mathbf{I}_{n_i}$ and D *unstructured*:

$$\boldsymbol{D} = \begin{bmatrix} \sigma^2_0 & \sigma_{01} & \sigma_{02} & \dots & \sigma_{0q} \\ \sigma_{10} & \sigma^2_1 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{20} & \sigma_{21} & \sigma^2_2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \sigma_{q2} & \dots & \sigma^2_q \end{bmatrix}$$

Inference for LMMs

- Hence, the vector of variance-covariance parameters is $\boldsymbol{\alpha} = \left[\sigma_\epsilon^2, \boldsymbol{D}\right]$

- In the frequentist domain, the random part component is used to model within-individual correlations, serial correlations, etc.

- Frequentist inferences usually focuses primarily on the fixed effects regression parameters $\boldsymbol{\beta}$ and the variance components α, although we could also potentially be interested on the random effects $\boldsymbol{b} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^{\mathrm{T}}$.

- That is, we are often interested in the marginal and mean variances:

$$\mathrm{E}\left[\boldsymbol{Y}_i|\boldsymbol{\beta}\right] = \boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta}$$

$$\mathrm{var}\left(\boldsymbol{Y}_i|\boldsymbol{\alpha}\right) = \boldsymbol{V}_i(\boldsymbol{\alpha}) = \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}} + \sigma_\epsilon^2\mathbf{I}_{n_i}$$

Inference of LMMs

- In order to conduct inference on the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we need to integrate over the random effects in the two-stage model:

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \int_{\mathcal{S}_{\boldsymbol{b}}} p(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\alpha}) \times p(\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{\alpha}) d\boldsymbol{b}$$

- Due to the conditional independencies in the likelihood

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{i=1}^{m} \int_{\mathcal{S}_{\boldsymbol{b}_i}} p\left(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\beta},\sigma_\epsilon^2\right) \times p\left(\boldsymbol{b}_i|\boldsymbol{D}\right) d\boldsymbol{b}_i$$

  and since a convolution of normal distributions is still normal, we have

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\alpha} \sim \mathbf{N}_{n_i}\left[\boldsymbol{\mu}_i(\boldsymbol{\beta}),\boldsymbol{V}_i(\boldsymbol{\alpha})\right]$$

  .

- The log-likelihood is

$$(\boldsymbol{\beta},\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{V}_i(\boldsymbol{\alpha})| - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\,\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}\,(\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$$

- We need to maximize the previous expression with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We obtain the score functions:

$$\begin{aligned}\frac{\partial l}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i - \sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\boldsymbol{x}_i\boldsymbol{\beta} \\ &= \sum_{i=1}^{m}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{V}_i^{-1}\,(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) \\ \frac{\partial l}{\partial\alpha_r} &= \frac{1}{2}\left\{(y - \boldsymbol{x}_i\beta)'\boldsymbol{V}_i^{-1}\frac{\partial\boldsymbol{V}_i}{\partial\alpha_r}\boldsymbol{V}_i^{-1}(y - \boldsymbol{x}_i\beta) - \mathrm{tr}\left(\boldsymbol{V}_i^{-1}\frac{\partial\boldsymbol{V}_i}{\partial\alpha_r}\right)\right\}\end{aligned}$$

$r = 1,\ldots,q$

- Hence, we obtain the MLE for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i\right)$$

which is a generalized least squares estimator (GLS).

- Hence, we obtain the MLE for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i \right)$$

  which is a generalized least squares estimator (GLS).

- $\boldsymbol{V}_i(\hat{\boldsymbol{\alpha}})$ is a function of an estimate $\hat{\alpha}$ of the variance components (plug-in).

- If $D = \mathbf{0}$, then $\boldsymbol{V} = \sigma_\epsilon^2 \mathbf{I}_N$, $N = \sum_{i=1}^m n_i$, and $\widehat{\boldsymbol{\beta}}$ corresponds to the ordinary least squares estimator

- It is easy to prove that

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\boldsymbol{\alpha})^{-1} \boldsymbol{x}_i \right)^{-1}$$

Some remarks

- The result can be seen as the maximization of a profile likelihood of $\boldsymbol{\beta}$ given $V(\boldsymbol{\alpha})$ or an estimate thereof.
- $\boldsymbol{V}_i(\hat{\boldsymbol{\alpha}}) = \boldsymbol{Z}_i D(\hat{\boldsymbol{\alpha}})\boldsymbol{Z}_i' + \boldsymbol{E}_i$
- The expected information matrix is block diagonal:

$$\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix}$$

  so there is asymptotic independence between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ and any consistent estimator of α will give an asymptotically efficient estimator for β
- The estimator $\widehat{\boldsymbol{\beta}}$ is linear in the data Y_i, and so under normality of the data, $\hat{\boldsymbol{\beta}}$ is normal also.

- Under correct specification of the variance model, and with a consistent estimator $\widehat{\boldsymbol{\alpha}}$

$$\left(\sum_{i=1}^{m} x_i V_i(\widehat{\alpha})^{-1} x_i\right)^{1/2} \left(\widehat{\beta}_m - \beta\right) \rightarrow_d \mathrm{N}_{k+1}(0, \mathrm{I})$$

  as $m \rightarrow \infty$.

- Since $\widehat{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$, These properties are valid for large samples even if the sampling distribution of Y_i is not multivariate normal (Laird and Ware, 1982).
- The second moments of the data need to be correctly specified.

Optimization

- The detailed maximization process in linear mixed models follows the standard procedures applied for general linear models.
- More specifically, optimization of the profiled log-likelihood is usually accomplished through EM iterations or through Newton-Raphson iterations (Laird and Ware, 1982)
- The EM algorithm (Dempster, Laird and Rubin, 1977) is a popular iterative algorithm for likelihood estimation in models with incomplete data.
- The EM iterations for the LME model are based on regarding the random effects, such as the $\boldsymbol{b}_i, i = 1, \ldots, M$, as unobserved data.
- Each iteration of the EM algorithm results in an increase in the log-likelihood, till convergence (stats 230)

EM algorithm - idea

- Starting from parameter guesses, $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the following steps are iterated:

  1. Find the distribution of $\boldsymbol{b}|\boldsymbol{y}$ according to the current parameter estimates (we will see soon how to compute $\boldsymbol{b}|\boldsymbol{y}$)

  2. Treating the distribution from 1 as fixed (rather than depending on $\boldsymbol{\alpha}, \boldsymbol{\beta}$), find an expression for $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{E}_{|\boldsymbol{y}}\{\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})\}$ as a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, using the distribution from 1. The $\mathbf{y}$ are treated as fixed, here. (This is the E-step.)

  3. Maximize the expression for $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t. the parameters to obtain updated estimates $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$. (This is the M-step.)

EM algorithm - idea

- Note that the expectation in step 2 is taken with respect to the fixed distribution from step 1, which depends on the current parameter estimates. When evaluating $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we view $\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$ as a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, but do not treat the distribution of $\mathbf{b}|\mathbf{y}$ as depending on these parameters.

Newton-Raphson

- The Newton-Raphson algorithm (Thisted, 1988) is one of the most widely used optimization procedures.
- It uses a first-order expansion of the score function (the gradient of the log-likelihood function) around the current estimate $\boldsymbol{\alpha}^{(w)}$ to produce the next estimate $\boldsymbol{\alpha}^{(w+1)}$.
- Each Newton-Raphson iteration requires the calculation of the score function and its derivative, the Hessian matrix of the log-likelihood.
- Under general conditions usually satisfied in practice, the Newton-Raphson algorithm converges quadratically.
- Because the calculation of the Hessian matrix at each iteration may be computationally expensive, simple, quicker to compute approximations are sometimes used, leading to the so-called Quasi-Newton algorithms.

Optimization

- Individual iterations of the EM algorithm are quickly and easily computed.
- Although the EM iterations generally bring the parameters into the region of the optimum very quickly, progress toward the optimum tends to be slow when near the optimum.
- Newton-Raphson iterations, on the other hand, are individually more computationally intensive than the EM iterations, and they can be quite unstable when far from the optimum. However, close to the optimum they converge very quickly
- A hybrid approach starts with an initial $\boldsymbol{\alpha}^{(0)}$, performing a moderate number of EM iterations, then switches to Newton-Raphson iterations.
- The lme function in the nlme package of R implements such a hybrid optimization scheme. It begins by calculating initial estimates of the $\boldsymbol{\alpha}$ parameters, then uses several EM iterations to get near the optimum, then switches toNewton-Raphson iterations to complete the convergence to the optimum. By default 25 EM iterations are performed before switching to Newton-Raphson iterations.

Testing Fixed effects

- Tests of fixed effects are typically done with either Wald or likelihood ratio (LRT) tests. With asymptotic distributions and independent predictors, Wald and LRT tests are equivalent.
- When a data set size is not large enough to be a good approximation of the asymptotic distribution or there is some correlation amongst the predictors, the Wald and LRT test results can vary considerably.
- Let $\tilde{\boldsymbol{L}}$ be a design vector or a design matrix of known weights for selected components in $\boldsymbol{\beta}$ and $\boldsymbol{L}\boldsymbol{\beta}$ be a combination of interest.
- The sampling distribution of $\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{L}\boldsymbol{\beta}$ and covariance matrix

$$\begin{aligned}\operatorname{cov}(\boldsymbol{L}\hat{\boldsymbol{\beta}}) &= \tilde{\boldsymbol{L}}\operatorname{cov}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{L}}' \\ &= \tilde{\boldsymbol{L}}\left[\left(\sum_{i=1}^{N}\boldsymbol{X}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i\right)^{-1}\right]\tilde{\boldsymbol{L}}'\end{aligned}$$

- Empirically, $\tilde{\boldsymbol{L}}$ is often designed to contain weight 1 to indicate the selected components in β or weight 0 for the components not selected.

Wald test

- The two hypotheses, $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\boldsymbol{\beta} \neq 0$ can be tested by using the following Wald statistic:

$$W^2 = (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})$$

where W^2 is the Wald statistic that asymptotically follows a chi-square distribution with rank($\tilde{\boldsymbol{L}}$) as the degrees of freedom.

Wald test

- The two hypotheses, $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\boldsymbol{\beta} \neq 0$ can be tested by using the following Wald statistic:

$$W^2 = (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})$$

  where W^2 is the Wald statistic that asymptotically follows a chi-square distribution with rank$(\tilde{\boldsymbol{L}})$ as the degrees of freedom.

- Similarly, the approximate confidence interval, given α, is given by

$$\tilde{L}\hat{\beta} \pm t_{\mathrm{df},\alpha/2} \times \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X} \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{\frac{1}{2}} .$$

- For a test of a single component, a Z-test works as an approximate Wald test.

Bias of the Wald statistics

- The Wald statistics is considered to be biased downward because the variability in estimating the variance components is not considered (Dempster et al. 1981) in the ML estimates.
- It is perceived that this bias can be resolved by using approximate F-statistic about β.
- Let $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\beta = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\beta \neq 0$, then we can use the F -statistic:

$$F = \frac{(\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1}(\hat{\alpha}) \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})}{\mathrm{rank}(\tilde{\boldsymbol{L}})}$$

where the degrees of freedom for the numerator is $\mathrm{rank}(\boldsymbol{L})$ and the degrees of freedom for the denominator needs to be estimated from the data. The uncertainty about the degrees of freedom for the denominator somewhat restricts the use of the F -test (several methods, omit).

Likelihood ratio test

- To test $\mathrm{H}_{0:} : \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_{\mathrm{A:}} : \tilde{\boldsymbol{L}}\beta \neq 0$ we could use the LRT.
- The likelihood ratio test compares the maximized log-likelihoods between two models, given by

  $$G^2 = 2\log L\left(\hat{\theta}_{\text{full}}\right) - 2\log L\left(\hat{\theta}_{\text{reduced}}\right)$$

  where G^2 is the likelihood ratio statistic, $\log L\left(\hat{\theta}_{\text{reduced}}\right)$ is the log-likelihood function for the model without one or more parameters, and $\log L\left(\hat{\theta}_{\text{full}}\right)$ is the log-likelihood function containing all parameters.
- The likelihood ratio statistic is asymptotically distributed as χ^2 with the degrees of freedom being the difference in the number of fixed-effects parameters.
- If G^2 is associated with a p -value smaller than α, the null hypothesis about θ should be rejected.