Statistical Methods for Correlated Data

Inference for Variance Components

Michele Guindani

Department of Biostatistics
UCLA

Inference for Variance Components via Maximum Likelihood

- The MLE $\widehat{\boldsymbol{\alpha}}$ is obtained from maximization of the log-likelihood but, in general, there is no closed form solution.
- Under some regularity conditions, the distribution of the ML (as well as REML estimator) $\widehat{\boldsymbol{\alpha}}$ can be well approximated by a normal distribution with mean vector α and with covariance matrix given by the inverse of the Fisher information matrix:

$$\boldsymbol{I}_{\alpha\alpha}^{1/2}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \rightarrow_d \mathbf{N}_r\left(\mathbf{0}, \mathbf{I}_r\right)$$

  where r is the number of distinct elements of $\boldsymbol{\alpha}$. One can compute CIs based on this asymptotic result.

Inference for Variance Components via Maximum Likelihood

- Using the asymptotic normality of the parameter estimates, approximate Wald tests and approximate Wald confidence intervals can now easily be obtained. However, the performance of the normal approximation strongly depends on the true value $\boldsymbol{\alpha}$, with larger samples needed for values of $\boldsymbol{\alpha}$ relatively closer to the boundary of the parameter space of $\boldsymbol{\alpha}$. In the case that $\boldsymbol{\alpha}$ is a boundary value (e.g. $\sigma_b^2 = 0$ in a random intercept model), the normal approximation completely fails.

Inference for Variance Components via Maximum Likelihood

- Similarly, the regularity conditions under which the chi-squared approximation for the LRT is valid include that H_0 should not define an hypothesis on the boundary of the parameter space of $\boldsymbol{\alpha}$.
- Stram and Lee (1994, 1995) have been able to show that the asymptotic null distribution for the likelihood ratio test statistic for testing hypotheses at the boundary of the parameter space of $\boldsymbol{\alpha}$ is often a mixture of chi-squared distributions rather than the classical single chi-squared distribution.
- By using the typical chi-squared test, one would fail to reject the null too often, leading to simpler covariance structures.

Case 1: No random effects versus one random effect

- We start by considering the random intercept model

$$Y_{ij} = \beta_0 + \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \epsilon_{ij}$$

with $b_i|\sigma_0^2 \sim \mathbf{N}\left(0, \sigma_0^2\right)$. We wish to test whether the random effects variance is zero, that is, $H_0 : \sigma_0^2 = 0$ versus $H_1 : \sigma_0^2 > 0$, i.e.:

$$H_0 : D = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{vs} \quad H_1 : D = \begin{pmatrix} d_{11} = \sigma_0^2 & 0 \\ 0 & 0 \end{pmatrix}$$

- In this case, the asymptotic null distribution is a 50: 50 mixture of χ_0^2 and χ_1^2 distributions, where the former is the distribution that gives probability mass 1 to the value 0.

Case 2: One versus two random effects (e.g. random slope & intercept)

- In the case one would like to test

$$H_0 : D = \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{vs} \quad H_1 : D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

  the asymptotic null distribution of the LRT is a mixture with equal weights 0.5 for χ^2_2 and χ^2_1.

CASE 3: K versus K+1 random effects

- For testing the hypothesis

$$H_0 : D = \begin{pmatrix} D_{11} & 0 \\ 0' & 0 \end{pmatrix}$$

  with D_{11} a $K \times K$ positive definite matrix vs
  H_1 : D is a general $(K+1) \times (K+1)$ positive semidefinite matrix, the large-sample behavior of the LRT is a mixture of χ^2_{k+1} and χ^2_k again with equal weights 0.5.

- The results assume that the likelihood function can be maximized over the space Θ_α of positive semidefinite matrices D, and that the estimating procedure is able to converge, for example, to values of D which are positive semidefinite but not positive definite. This is software dependent and should be checked when the above results are applied in practice.

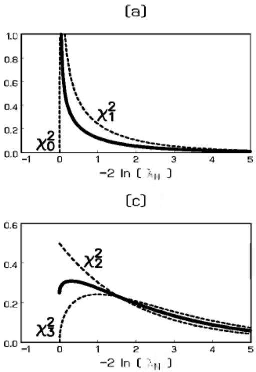

FIGURE 6.3. *Graphical representation of the asymptotic null distribution of the likelihood ratio statistic for testing the significance of random effects in a linear mixed model, for three different types of hypotheses. For each case, the distribution (solid line) is a mixture of two chi-squared distributions (dashed lines), with both weights equal to 0.5:*
(a) Case 1: no random effects versus one random effect.
(b) Case 2: one random effect versus two random effects.
(c) Case 3: two random effects versus three random effects.

General Result

In general, Shapiro showed that when testing jointly K parameters, the mixture is a complex mixture of χ^2_m,

$$\sum_{m=0}^{K} 2^{-K} \binom{K}{m} \chi^2_m$$

Statistical Methods for Correlated Data

Inference for Variance Components via Restricted Maximum Likelihood (REML) inference

Michele Guindani

Department of Biostatistics
UCLA

- The MLE of the variance components is generally biased

- The MLE of the variance components is generally biased
- This is well-known in typical linear regression:

$$\boldsymbol{Y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I_n)$$

- The MLE estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\boldsymbol{\beta}})'\,(Y - X\hat{\boldsymbol{\beta}})}{n} = \frac{RSS}{n}$$

with RSS=residual sum of squares, $E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2$.

whereas the usual unbiased estimator is

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

with p=number of parameters in $\boldsymbol{\beta}$.

- The ML estimates of variance components are biased:
  - ☞ They do not take into account that $\boldsymbol{\beta}$ is also estimated
  - ☞ The variances are then underestimated
  - ☞ 95% CI are shorter
  - ☞ biased OCs of test procedures

Restricted likelihood (REML) estimation

- Basic Idea: separate data in two parts: one used for estimation of $\boldsymbol{\beta}$ and one used for estimating $\boldsymbol{\alpha}$

☞ The estimation of $\boldsymbol{\alpha}$ is based on a profile likelihood that does not depend on $\boldsymbol{\beta}$

☞ A possible way to do this is to consider the residuals after estimating $\boldsymbol{\beta}$ via OLS

☞ Similarly, one could obtain a marginal likelihood of $\boldsymbol{\alpha}$ that does not depend on $\boldsymbol{\alpha}$.

Error Contrasts

- To derive a marginal likelihood, we need to find a function of the data that $\boldsymbol{U} = f(\boldsymbol{Y})$ that does not depend on $\boldsymbol{\beta}$.

Error Contrasts

- To derive a marginal likelihood, we need to find a function of the data that $\boldsymbol{U} = f(\boldsymbol{Y})$ that does not depend on $\boldsymbol{\beta}$.
- We define an error contrast a linear function $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}$ such that

$$\mathrm{E}\left[\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}\right] = 0 \text{ for all values of } \boldsymbol{\beta}$$

  with C an N -dimensional vector.
- It is easy to show that for a LMM:

$$\mathrm{E}\left[\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Y}\right] = 0 \text{ for all } \boldsymbol{\beta} \text{ if and only if } \boldsymbol{C}^{\mathrm{T}}\boldsymbol{x} = 0$$

Error Contrasts

- Then, when $C^{\mathrm{T}}x = 0$

$$\boldsymbol{C}^T\boldsymbol{Y} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{z}\boldsymbol{b} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\epsilon}$$

  which does not depend on β, suggesting that the marginal likelihood could be based on error contrasts.

Error Contrasts

- Then, when $C^{\mathrm{T}}x = 0$

$$\boldsymbol{C}^{T}\boldsymbol{Y} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{z}\boldsymbol{b} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\epsilon}$$

  which does not depend on β, suggesting that the marginal likelihood could be based on error contrasts.

- Let's consider the matrix such that

$$\boldsymbol{B}\,\boldsymbol{B}^{T} = I - \boldsymbol{x}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}$$

  which converts $\boldsymbol{Y}$ to OLS residuals

- Then

$$\boldsymbol{U} = \boldsymbol{B}^{T}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}(\mathbf{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{r}$$

  with $\boldsymbol{H} = \boldsymbol{x}\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}$ and $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{o}}$, and $\widehat{\boldsymbol{\beta}}_{\mathrm{o}} = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y}$ (OLS estimator). Easy to check $E(\boldsymbol{U}) = 0$.

REML estimator

- We need to obtain the distribution of $\boldsymbol{U}$ and show that it does not depend on $\boldsymbol{\beta}$.

REML estimator

- We need to obtain the distribution of $\boldsymbol{U}$ and show that it does not depend on $\boldsymbol{\beta}$.
- We consider the transformation

$$\boldsymbol{Y} \to \left[\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\mathrm{G}}\right] = \left[\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}, \boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y}\right]$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{G}} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{Y} = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{Y}$ is GLS estimator

- The vector $\left[\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{\mathrm{c}}\right]$ is a linear combination of normals and so is normal, and

$$\begin{aligned}
\mathrm{cov}\left(\boldsymbol{U}, \widehat{\boldsymbol{\beta}}_{G}\right) &= \mathrm{E}\left[\boldsymbol{U}\left(\widehat{\boldsymbol{\beta}}_{G} - \boldsymbol{\beta}\right)^{\mathrm{T}}\right] \\
&= \mathrm{E}\left[\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{G}\right] - \mathrm{E}\left[\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\beta}^{\mathrm{T}}\right] \\
&= \boldsymbol{B}^{\mathrm{T}}\left[\mathrm{var}(\boldsymbol{Y}) + \mathrm{E}(\boldsymbol{Y})\mathrm{E}\left(\boldsymbol{Y}^{\mathrm{T}}\right)\right]\boldsymbol{G} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{\beta}\boldsymbol{\beta}^{\mathrm{T}} \\
&= \boldsymbol{B}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{G} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}\boldsymbol{\beta}(\boldsymbol{x}\boldsymbol{\beta})^{\mathrm{T}} \\
&= \boldsymbol{0}
\end{aligned}$$

➽ $\boldsymbol{U}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{G}}$ are uncorrelated and, since they are normal, also independent.

REML estimator

- Then,

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{\alpha},\boldsymbol{\beta}) &= p\left(\boldsymbol{U},\widehat{\boldsymbol{\beta}}_{\mathrm{G}}|\boldsymbol{\alpha},\boldsymbol{\beta}\right)|\boldsymbol{J}| \\ &= p(\boldsymbol{U}|\boldsymbol{\alpha})p\left(\widehat{\boldsymbol{\beta}}_{\mathrm{G}}|\boldsymbol{\alpha},\boldsymbol{\beta}\right)|\boldsymbol{J}| \end{aligned}$$

- where

$$p(\boldsymbol{U}|\boldsymbol{\alpha}) = c\frac{\left|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right|^{1/2}|\boldsymbol{V}|^{-1/2}}{\left|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{x}\right|^{1/2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}\right)^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}\right)\right]$$

with $c = (2\pi)^{-(N-k-1)/2}$, that does not depend upon the transformation $\boldsymbol{B}$. So, one could choose any independent combination of the residuals

REML estimator

- The restricted log-likelihood upon which base inference for $\boldsymbol{\alpha}$

$$l_R(\boldsymbol{\alpha}) = -\frac{1}{2}\log\left|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\alpha})^{-1}\boldsymbol{x}\right| - \frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\alpha})| - \frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}\right)^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\alpha})^{-1}\left(\boldsymbol{y}-\boldsymbol{x}\widehat{\boldsymbol{\beta}}_{\mathrm{G}}\right)$$

  with the extra-term $(= -\log(Cov(\hat{\beta}))$ highlighted w.r.t. to ML.
- Use EM or/and Newton-Raphson for optimization

When to use ML and REML

- Many authors have discussed relative merits of ML and REML estimation for covariance matrices.
- They behave similarly for n (# of occasions) and m (# of individuals) large, with p fixed
- When $p \uparrow$, comparisons unequivocally favor REML
- General suggestion: use the REML estimator for Σ_i (less seriously biased). When the REML is used to estimate Σ_i, $\hat{\beta}$ is usually estimated as

$$\hat{\boldsymbol{\beta}}_{\mathrm{G}} = \left(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{x}\right)^{-1} \boldsymbol{x}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{Y}$$

using $V_i(\hat{\boldsymbol{\alpha}})_{REML}$ as a plug-in estimate

However...

However......for LRT testing

- The REM log-likelihood can be used to compare nested models for the covariance, but should not be used to compare regression models characterized by different fixed effects:
  - The extra-determinant $-\frac{1}{2}\log|\boldsymbol{x}^T V^{-1}\boldsymbol{x}|$ depends on X (which changes across two models with different fixed effects)
  - The transformation $Y = \boldsymbol{B}^T = B^T\boldsymbol{r}$ depends on X through the residuals ☞ hence, the two models will not be nested for $\hat{\beta}$
- Instead, the standard ML log-likelihood should be used for constructing LRTs that compare nested regression models for the mean

Summary

- REML:
  - Estimation of fixed effects and testing with Wald test (see Example Dental Growth Curves, Sect. 8.5.3)
  - Variance-covariance estimation
  - Inference & Prediction of random effects in LME
  - LRT comparing nested covariance or random effects specifications
- ML:
  - Estimation and Inference of fixed effects
  - LRT comparing models differing only in the fixed effect structure
  - Testing with Wald test