Linear Mixed Models

- Many modeling techniques assume the data are independent and identically distributed (i.i.d.), this assumption does not hold in many cases
- Linear mixed models (LMM) are used to model dependent data where the error structure is defined in the model
- LMM is primarily used with data that are grouped according to one or more naturally occurring classification factors
- LMMs combine both fixed and random effects to handle the dependent structure in the data
- In econometrics, often referred to as a random-coefficient regression model; in the social sciences, it is often called a multilevel or hierarchical linear model

A general linear mixed model may be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{1}$$

- $\mathbf{Y}$ is an N-dimensional response vector
- $\mathbf{X}$ and $\mathbf{Z}$ are known $N \times p$ and $N \times q$ matrices of covariates, respectively
- $\boldsymbol{\beta}_{p\times 1}$ is a vector of unknown regression coefficients, which are often called the fixed effects, $\boldsymbol{\alpha}_{q\times 1}$ is a vector of random effects and $\boldsymbol{\epsilon}_{N\times 1}$ is a vector of errors
- Basic assumptions:

  —-The random effects and errors have mean zero and finite variances. Typically, the covariance matrices $\mathbf{G} = Var(\boldsymbol{\alpha})$ and $\mathbf{R} = Var(\boldsymbol{\epsilon})$ involve some unknown dispersion parameters, or variance components

  —-The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ are assumed to be uncorrelated

- A fixed effect is not subject to variation and is considered to be an unknown constant, such as an effect of a treatment
- A random effect is a factor level randomly selected from the collection of all possible factor levels, such as randomly selected subjects in an experiment. Random effects are usually not of interest
- Mixed effects data contain at least one random effect and some fixed effects
- Assume that observations that share the same combination of classification factors also share the same corresponding random effect, while all observations share the fixed effects
- Some examples that deal with mixed effects are repeated measures data, longitudinal data, nested data, block designs and small area estimation

Example 1: Nested Designs.

- Goal: to investigate the variation of emergency room treatment of patients in a given city

- Data: First, m number of hospitals are randomly selected from the given city; Then, n_i doctors $(i = 1, 2, \cdots, m)$ are randomly selected from each selected hospital; During a pre-specified period of time, the emergency care of patients treated by selected doctors is evaluated; An overall performance measure can be used that summarizes the treatment of ER patients from each selected doctor

- Data is nested: Impossible to force each doctor to work in each selected hospital, so data is not crossed but nested.

- Random effect: Each hospital represents a random effect since we are not uniquely interested in the hospitals selected but in the set of all hospitals in the city

- Dependency of the data: doctors within the same hospital tend to be more similar than doctors in different hospitals, so observations are not independent

Nested Error Regression Model

$$Y_{ij} = u + v_i + \epsilon_{ij} \tag{2}$$

- y_{ij} denote the measure of emergency room care from the jth doctor $(j = 1, 2, \cdots, n_i)$ from the ith hospital $(i = 1, 2, \cdots, m)$
- u is the overall true mean; $v_i \sim N(0, \sigma_v^2)$ is the random effect corresponding to the hospital; ϵ_{ij} are independent and assumed to be distributed $N(0, \sigma^2)$
- Testing whether the emergency room care differs significantly between hospitals is reduce to testing whether $\sigma_v^2 = 0$.

Example 2: Linear Mixed Model

- Data: The National Assessment of Educational Progress (NAEP). A wealth of information is collected for each student, teacher and participating school. School-level information includes fiscal resources, instructional methods, student-body characteristics, and expectations of academic achievement.
- Goal: Model relationship between mathematics proficiency score of student and gender of the student
- Y_{ij} be the mathematics proficiency score of student j at school i in the sample; $x_{ij} = 1$ if student j at school i is female and 0 if student j at school i is male

Consider model 1,

$$Y_{ij} = \beta_0 + x_{ij}\beta_1 + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2) \tag{3}$$

- Characteristics of the schools, teachers, and neighborhoods that are not included in the model may induce a positive correlation in the test scores within a school. For instance, the seventh and eighth mathematics teacher in one school might be superb at inspiring students to learn mathematics. The students from that school might then all perform better than average on the proficiency test, so their scores are more similar.

- Some schools may encourage students of one gender more than students of the other gender

- The model doesn’t allow for different relations between gender and test score in different schools
- The p-value for parameter estimates will be far too small
- This model is likely to be inappropriate

Consider a model that incorporates cluster effects and allows schools to have different slopes for gender

$$Y_{ij} = \beta_{0i} + (x_{ij} - \bar{x}_i)\beta_{1i} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

- β_{0i} can be interpreted as the average test score in school i
- School i has its own straight-line regression model with intercept β_{0i} and slope β_{1i}

Consider a model where slopes and intercepts from different schools are also related through a model

$$Y_{ij} = \beta_0 + (x_{ij} - \bar{x}_i)\beta_1 + \delta_{0i} + (x_{ij} - \bar{x}_i)\delta_{1i} + \epsilon_{ij} \quad (4)$$

- $\beta_{0i} = \beta_0 + \delta_{0i}; \beta_{1i} = \beta_1 + \delta_{1i}$ where δ_{0i} and δ_{1i} follow a bivariate normal distribution with $E_M[\delta_{0i}] = E_M[\delta_{1i}] = 0, V_M[\delta_{0i}] = \tau_{00}, V_M[\delta_{1i}] = \tau_{11}$, and $Cov_M(\delta_{0i}, \delta_{1i}) = \tau_{01}$
- β_0 represents the mean test score for schools
- β_1 represents the mean slope for gender for schools
- Random effects δ_{0i} and δ_{1i} represent the difference in the intercept and slope between school i and the average values for

intercept and slope for all schools; they measure the school effects

- ϵ_{ij} refers to additional deviation from the mean due to the individual student, after the effect of gender and school have been accounted for
- If $\tau_{00} = \tau_{11} = 0$, there is no school effect on test score, and the model then reduces to a regular straight-line regression model

Example 3: Small Area Estimation of Income

Small Area Estimation

- Large scale sample surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas

- There are smaller geographic areas and subpopulations for which adequate samples are not available

- It is necessary to “borrow strength” from related small areas to find indirect estimators that increase the effective sample size and thus precision. Such indirect estimators are typically base on linear mixed models or generalized linear mixed models that provide a link to related small area through the use of supplementary data such as recent census data and current administrative records

Example Fay and Herriot (1997)

- Use a linear mixed model for estimating per-cpita income (PCI) for small places from the 1970 Census of Population and Housing
- Income was collected on the basis of a 20 percent sample. However, of the estimates required, more than one-third, or approximately 15,000, were for places with population of fewer than 500 persons.
- With such small populations, the sampling error of the direct estimates is quite significant

- Need to borrow strength from related places and other sources

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i$$

where y_i is the natural logarithm of the sample estimate of PCI for the ith place (the logarithm transformation stabilized the variance); $\mathbf{x}_i$ is a vector of known covariates related to the place; $\boldsymbol{\beta}$ is a vector of unknown regression coefficients; v_i is a random effect associated with the place; and e_i represents the sampling error. Assume that v_i and e_i are distributed independently such that $v_i \sim N(0, \mathbf{A}), e_i \sim N(0, \mathbf{D}_i)$, where $\mathbf{A}$ is unknown but $\mathbf{D}_i$'s are known

Types of linear mixed models

- Gaussian linear mixed models (normality assumption)

- Non-Gaussian linear mixed models (without normality assumption)

Example 1: One-way random effects model

$$y_{ij} = u + \alpha_i + \epsilon_{ij} \tag{5}$$

where

- $i = 1, \cdots, m, j = 1, \cdots, k_i, \sum(k_i) = n$
- u is the overall true mean
- $\alpha_i \sim N(0, \sigma^2), \epsilon_{ij} \sim N(0, \tau^2)$
- the random effects are independent with the errors

Note:

- A model is called a random effects model if the only fixed effect is an unknown mean
- Typically, the variance σ^2 and τ^2 are unknown
- Model can be expressed in form (1)

Kronecker Product: $\mathbf{A}$ is an $m \times n$ matrix, $\mathbf{B}$ is an $p \times q$ matrix, then $\mathbf{A} \bigotimes \mathbf{B}$ is an $mp \times nq$ matrix

$$\mathbf{A} \bigotimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}$$

Let $\mathbf{y}_i = (y_{ij})_{1\le j\le k_i}$ be the column vector of observations from the ith group or cluster, and similarly $\boldsymbol{\epsilon}_i = (\epsilon_{ij})_{1\le j\le k_i}$. So $\mathbf{y} = (\mathbf{y}_1', \cdots, \mathbf{y}_m')'$, $\boldsymbol{\alpha} = (\alpha_i)_{1\le i\le m}$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \cdots, \boldsymbol{\epsilon}_m')'$ and $n = \sum k_i$. Consider $k_i = k$,

(5) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

where $\mathbf{X} = 1_m \bigotimes 1_k = 1_{mk}$, $\boldsymbol{\beta} = u$, $\mathbf{Z} = \mathbf{I}_m \bigotimes 1_k$, $\boldsymbol{\alpha} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_m)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tau^2\mathbf{I}_n)$

A general Gaussian linear mixed model may be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{6}$$

- $\mathbf{Y}$ is an N-dimensional response vector
- $\mathbf{X}$ and $\mathbf{Z}$ are known $N \times p$ and $N \times q$ matrices of covariates, respectively
- $\boldsymbol{\beta}_{p\times 1}$ is a vector of unknown regression coefficients, which are often called the fixed effects, $\boldsymbol{\alpha}_{q\times 1}$ is a vector of random effects and $\boldsymbol{\epsilon}_{N\times 1}$ is a vector of errors
- $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$
- The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$ are assumed to be independent
- $Var(\mathbf{Y}) = \mathbf{R}+\mathbf{Z}\mathbf{G}\mathbf{Z}' = \mathbf{V}$, Under normality, we have $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$

A compromise between Bayesian and frequentist approaches

Bayesian approach

- The model is specified in hierachical fashion as

$$\mathbf{y}|\boldsymbol{\theta} \sim L(\mathbf{y}|\boldsymbol{\theta}), \tag{7}$$

$$\boldsymbol{\theta} \sim G(\boldsymbol{\theta}) \tag{8}$$

- Equation (7) defines the conditional distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$ through density L. Equation (8) defines a member of a family of distribution of $\boldsymbol{\theta}$ through the distribution G

- The parameter specifies G is called the hyperparameter

- Unlike the frequentist approach, the bayesian approach assumes that parameter $\boldsymbol{\theta}$ is random and densities L and G must be specified completely

- Posterior density

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta})G(\boldsymbol{\theta})}{\int L(\mathbf{y}|\boldsymbol{\theta})G(\boldsymbol{\theta})}$$

Mixed model approach

- The model is specified as a hierachical model (7) and (8), but it is allowed to have nonrandom parameters $\boldsymbol{\beta}$

$$\mathbf{y}|\boldsymbol{\theta} \sim L(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}), \tag{9}$$

$$\boldsymbol{\theta} \sim G(\boldsymbol{\theta}, \boldsymbol{\beta}) \tag{10}$$

- $\boldsymbol{\beta}$ is known and is the hyperparameter

$$L(\boldsymbol{\beta}) = \int L(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\beta}) G(\boldsymbol{\theta}, \boldsymbol{\beta}) d\boldsymbol{\theta} \tag{11}$$

- Random effects are unobservable and are integrated out in (11), $\boldsymbol{\beta}$ is estimated

- We call $\boldsymbol{\theta}$ random and $\boldsymbol{\beta}$ fixed effects parameters
- The posterior mean is called the estimate of the random effect

Summary

- Mixed model = Bayesian + fequentist
- As in Bayesian approach, mixed model assumes a hierarchical model where the parameter is treated as random
- On the other hand, the hyperparameter, $\boldsymbol{\beta}$ is not arbitrarily specified as in the Bayesian approach, but is estimated from the data
- A mixed model is more flexible than the Bayesian approach

Estimation in Gaussian Models

- Maximum likelihood
- Restricted maximum likelihood

Maximum likelihood (Fisher 1922)

- used in mixed model analysis until Hartley and Rao (1967)
- The estimation of the variance components in a linear mixed model was not easy to handle computationally in the old days

Point estimation

- The distribution of $\mathbf{y}$ has a joint pdf

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{V}|^{1/2}} exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}$$

- Log-likelihood function is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = c - \frac{1}{2}log(|\mathbf{V}|) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

where $\boldsymbol{\theta}$ represents the vector of all the variance components (involved in $\mathbf{V}$), and c is a constant

- $$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \quad (12)$$

$$\frac{\partial l}{\partial \boldsymbol{\theta}_r} = \frac{1}{2}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}\frac{\partial V}{\partial \boldsymbol{\theta}_r}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - tr(\mathbf{V}^{-1}\frac{\partial V}{\partial \boldsymbol{\theta}_r})] \quad (13)$$

where $r = 1, 2, \cdots, q$, $\boldsymbol{\theta}_r$ is the rth component of $\boldsymbol{\theta}$, which has dimension q

- For simplicity, assume that $\mathbf{X}$ is of full column rank, rank$(\mathbf{X}) = p$, let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ be the MLE, from (12)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

where $\hat{\mathbf{V}} = V(\hat{\boldsymbol{\theta}})$, once the MLE of $\boldsymbol{\theta}$ is found, the MLE of $\boldsymbol{\beta}$ can be calculated by the “closed-form” expression

By (12) and (13),

$$\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}\mathbf{P}\mathbf{y} = tr(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}), r = 1, 2, \cdots, q$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

For one-way random effects model, the log-likelihood function is equivalent to

$$
\begin{aligned}
l(u, \sigma^2, \tau^2) &= c - \frac{1}{2}(n-m)log(\tau^2) \\
&- \frac{1}{2}\sum_{i=1}^{m} log(\tau^2 + k_i\sigma^2) \\
&- \frac{1}{2\tau^2}\sum_{i=1}^{m}\sum_{j=1}^{k_i}(y_{ij} - u)^2 \\
&+ \frac{\sigma^2}{2\tau^2}\sum_{i=1}^{m}\frac{k_i^2}{\tau^2 + k_i\sigma^2}(\bar{y}_{i.} - u)^2
\end{aligned}
$$

where $\bar{y}_{i.} = k_i^{-1} \sum_{j=1}^{k_i} y_{ij}$. We can find $\frac{\partial l}{\partial u}$, $\frac{\partial l}{\partial \tau^2}$ and $\frac{\partial l}{\partial \sigma^2}$ and set them equal to zero to find estimators $\hat{u}, \hat{\sigma}^2$ and $\hat{\tau}^2$.

Asymptotic covariance matrix

- Under suitable conditions, the MLE is consistent and asymptotically normal with the asymptotic covariance matrix equal to the inverse of the Fisher information matrix
- Let $\psi = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$, then, under regularity conditions, the Fisher information matrix has the following expressions

$$Var(\frac{\partial l}{\partial \psi}) = -E(\frac{\partial^2 l}{\partial \psi \partial \psi'})$$

Restricted Maximum likelihood (REML) (Thompson 1962, Patterson and Thompson 1971)

- Consider a transformation of the data that is orthogonal to the design matrix of the fixed effects
- uses a likelihood function calculated from the transformed set of data, so that nuisance parameters have no effect

  —Example: In a random effects model, the fixed effects are considered as nuisance parameters, while the main interest is the variance component

Example 1: (Neyman-Scott problem). Nayman and Scott (1948) gave the following example which shows that, when the number of parameters increases with the sample size, the MLE may not be consistent. Suppose that two observations are collected from m individuals. Each individual has its own (unknown) mean, say, u_i for the ith individual. Suppose that the observations are independent and normally distributed with variance σ^2. The problem of interest is to estimate σ^2.

$$y_{ij} = u_i + \epsilon_{ij}$$

where ϵ_{ij}'s are independent and distributed as $N(0, \sigma^2)$, $i = 1, 2, \cdots, m$.

- The number of parameters $m+1$ is proportional to $2m$, the number of observations of the data.
- MLE of σ^2 is inconsistent

$$\text{Let } \mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{m1} \\ y_{m2} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{m1} \\ \epsilon_{m2} \end{pmatrix}$$

$$\mathbf{X} = \mathbf{I}_m \bigotimes 1_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

$$
\begin{aligned}
\hat{\sigma}^2_{ML} &= \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{2m} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{2} (y_{ij} - \bar{y}_{i.})^2 / 2m \\
&= \sum_{i=1}^{m} \frac{1}{2} (y_{i1} - y_{i2})^2 / 2m \\
&= \frac{1}{2} \sum_{i=1}^{m} (y_{i1} - y_{i2})^2 / 2m
\end{aligned}
$$

Let $\mathbf{z}_i = y_{i1} - y_{i2}$, $\mathbf{z}_i \sim N(0, 2\sigma^2)$, $\mathbf{z}_i$'s iid,

$S^2 = \dfrac{\sum z_i^2}{m-1} \rightarrow_p 2\sigma^2$, Hence

$$
\begin{aligned}
\hat{\sigma}^2_{ML} &= \frac{1}{2} \cdot \frac{\sum z_i^2}{2m} \\
&= \frac{1}{2} \cdot \frac{\sum z_i^2}{m-1} \cdot \frac{m-1}{2m} \\
&\rightarrow_p \frac{1}{2} \cdot 2\sigma^2 \cdot \frac{1}{2} \\
&= \frac{\sigma^2}{2}
\end{aligned}
$$

- Consider transformation: $z_i = y_{i1} - y_{i2}$.

  It follows that $z_1, \cdots, z_m$ are independent and distributed as $N(0, 2\sigma^2)$. u_i's are gone.

  Let $\mathbf{z} = \mathbf{A}'\mathbf{y}$

$$\mathbf{A}' = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

  Notice $\mathbf{X} = \mathbf{I}_m \bigotimes 1_2$, $\mathbf{A}'\mathbf{X} = \mathbf{0}$.

- In this example, RMLE is to apply a transformation to the data to eliminate the fixed effects, then use the transformed data to estimate the variance component

- $\hat{\sigma}^2_{REML} = \frac{1}{m} \sum_{i=1}^{m} (z_i - \bar{z})^2 \rightarrow_p \sigma^2$

Point Estimation

- Without loss of generality, assume that rank(X) =p
- Let $\mathbf{A}$ be an $n \times (n-p)$ matrix such that rank$(\mathbf{A}) = n - p, \mathbf{A}'\mathbf{X} = \mathbf{0}$
- Define $\mathbf{z} = \mathbf{A}'\mathbf{y}$, then $\mathbf{z} \sim N(\mathbf{0}, \mathbf{A}'\mathbf{V}\mathbf{A})$
- The joint pdf of $\mathbf{z}$ is given by

$$f_R(\mathbf{z}) = \frac{1}{(2\pi)^{(n-p)/2}|\mathbf{A}'\mathbf{V}\mathbf{A}|^{1/2}} exp\left\{-\frac{1}{2}\mathbf{z}'(\mathbf{A}'\mathbf{V}\mathbf{A})^{-1}\mathbf{z}\right\}$$

where the subscript R corresponds to “restricted”

Restricted log-likelihood is given by

$$l_R(\boldsymbol{\theta}) = c - \frac{1}{2}\text{log}(|\mathbf{A}'\mathbf{V}\mathbf{A}|) - \frac{1}{2}\mathbf{z}'(\mathbf{A}'\mathbf{V}\mathbf{A})^{-1}\mathbf{z} \tag{14}$$

Differentiating the restricted log-likelihood, we obtain

$$\frac{\partial l_R}{\partial \boldsymbol{\theta}_i} = \frac{1}{2}\left\{\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_i}\mathbf{P}\mathbf{y} - tr\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_i}\right)\right\}, i = 1, \cdots, q \tag{15}$$

where

$$\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{V}\mathbf{A})^{-1}\mathbf{A}' \tag{16}$$

- The REML estimator of $\boldsymbol{\theta}$ is defined as the maximizer of (14)

Note:

- Although the REML estimator of $\boldsymbol{\theta}$ is defined through transforming matrix $\mathbf{A}$, the REML estimator doesn't depend on $\mathbf{A}$
- The choice of $\mathbf{A}$ is not unique, but the results will be the same
- The REML method is a method of estimating $\boldsymbol{\theta}$, (not $\boldsymbol{\beta}$, $\boldsymbol{\beta}$ is eliminated before the estimation)
- Once the REML estimator of $\boldsymbol{\theta}$ is obtained, $\boldsymbol{\beta}$ is usually estimated the same way as the ML. Such estimator is sometimes referred as the "REML estimator" of $\boldsymbol{\beta}$

Example: one-way balanced random effect model

$$y_{ij} = u + \alpha_i + \epsilon_{ij}$$

where

- $i = 1, \cdots, m, j = 1, \cdots, k, mk = n$
- u is the overall true mean
- $\alpha_i \sim N(0, \sigma^2), \epsilon_{ij} \sim N(0, \tau^2)$
- the random effects are independent with the errors

Let $SS_{total} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{..})^2$, $SSA = k\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2$

and $SSE = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{i.})^2$

The loglikelihood function (14) is equivalent to

$$\begin{aligned} l_R &= -\frac{1}{2}(mk-1)\text{log}2\pi - \frac{1}{2}\text{log}(mk) - \frac{1}{2}m(k-1)\text{log}\tau^2 \\ &- \frac{1}{2}(m-1)\text{log}(\tau^2 + k\sigma^2) - \frac{SSE}{2\tau^2} - \frac{SSA}{2(\tau^2 + k\sigma^2)} \end{aligned}$$

Set $\dfrac{\partial l}{\partial \sigma^2}$ and $\dfrac{\partial l}{\partial \tau^2}$ equal to zero respectively, we have

$$\tau^2 + k\sigma^2 = SSA/(m-1) = MSA$$

and

$$\tau^2 = \frac{SSE}{n-m} = MSE$$

The REML equations thus have an explicit solution $\hat{\tau}^2 = MSE$ and $\hat{\sigma}^2 = k^{-1}(MSA - MSE)$

Asymptotic covariance matrix

- under suitable conditions, the REML estimator is consistent and asymptotically normal. The asymptotic covariance matrix is equal to the inverse of the restricted Fisher information matrix. Under the regularity conditions,

$$Var\left(\frac{\partial l_R}{\partial \boldsymbol{\theta}}\right) = -E\left(\frac{\partial^2 l_R}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)$$

Assuming that V is twice condinuously differentiable (with respect to the components of q), then we have

$$E\left(\frac{\partial^2 l_R}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right) = -\frac{1}{2}\text{tr}\left(\mathbf{P}\frac{\partial V}{\partial \boldsymbol{\theta}_i}\mathbf{P}\frac{\partial V}{\partial \boldsymbol{\theta}_j}\right), 1 \leq i, j \leq q$$

Analysis of variance estimation

- The basic idea of ANOVA estimation came from the method of moments
- Let Q be q-dimentional vector whose components are quadratic functions of the data. The ANOVA estimators of the variance components are obtained by solving the system of equations $E(Q) = Q$

Example: One way balanced random effects model continued

ANOVA table (consider the α_i's are fixed treatment effects):

source	SS	df	MS	F
Treatment	SSA	m-1	MSA=SSA/(m-1)	MSA/MSE
Error	SSE	m(k-1)	MSE=SSE/m(k-1)	
Total	SS_{total}	mk-1		

$SS_{total} = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{..})^2$, $SSA = k\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2$ and

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \bar{y}_{i.})^2$$

- The components of Q consist of SSA and SSE

$$E(SSA) = (m-1)k\sigma^2 + (m-1)\tau^2$$

and

$$E(SSE) = m(k-1)\tau^2$$

- The ANOVA estimating equations are

$$(m-1)k\sigma^2 + (m-1)\tau^2 = SSA$$

$$m(k-1)\tau^2 = SSE$$

- The resulting ANOVA estimators are therefore $\hat{\sigma}^2 = (MSA - MSE)/k$, $\hat{\tau}^2 = MSE$.

Note:

- unlike ML and REML, ANOVA estimators of he variance components may not belong to the parameter space. If $MSA < MSE$, $\hat{\sigma}^2$ will be negative.
- Under a balanced mixed ANOVA model, the ANOVA estimator of $\boldsymbol{\theta} = (\tau^2, \sigma_i^2, 1 \leq i \leq s)'$ is identical to the solution of the REML equations.

—The solution of the REML equations is not necessrily the REML estimator, because the REML estimator has to be in the parameter space

—When the solution does belong to the parameter space, the REML and ANOVA estimators are identical

Unbalanced data (Hendenson (1953))

Tests in Gaussian Mixed Models

Some notations:

- Projection matrix $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- $\mathbf{Z} \ominus \mathbf{X} = P_{\mathbf{X}^{\perp}}\mathbf{Z}$
- $P_{\mathbf{X}^{\perp}} = \mathbf{I} - P_{\mathbf{X}}$

Exact F test

Suppose that one wishes to test the hypothesis $H_0 : \sigma_1^2 = 0$.

Note that the model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\alpha}_1 + \mathbf{Z}_{-1}\boldsymbol{\alpha}_{-1} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\alpha}_{-1} = (\boldsymbol{\alpha}_2', \cdots, \boldsymbol{\alpha}_s')'$, $\mathbf{Z}_{-1} = (\mathbf{Z}_2, \cdots, \mathbf{Z}_s)$

Let

$$\begin{aligned} q_1 &= \tau^{-2}\mathbf{y}' P_{\mathbf{Z}_1 \ominus (\mathbf{X}, \mathbf{Z}_{-1})}\mathbf{y} \\ &= \mathbf{y}' \left\{ P_{\mathbf{Z}_1 \ominus (\mathbf{X}, \mathbf{Z}_{-1})/\tau^2} \right\} \mathbf{y} \end{aligned}$$

Let

$$
\begin{aligned}
q_2 &= \tau^{-2}\mathbf{y}' P_{(\mathbf{X},\mathbf{Z})^\perp}\mathbf{y} \\
&= \mathbf{y}' \left\{P_{(\mathbf{X},\mathbf{Z})^\perp/\tau^2}\right\} \mathbf{y}
\end{aligned}
$$

Then $q_1 \sim \chi^2_{r_1}$ with $r_1 = rank\left\{(\mathbf{X},\mathbf{Z})\right\} - rank\left\{(\mathbf{X},\mathbf{Z}_{-1})\right\}$
$q_2 \sim \chi^2_{r_2}$ with $r_2 = n - rank\left\{(\mathbf{X},\mathbf{Z})\right\}$

- q_1 and q_2 are independent
- $F_1 = \dfrac{q_1/r_1}{q_2/r_2} \sim F_{r_1,r_2}$

Optimal Test (Mathew & Sinha (1988)) (Jiang, Chapter 2, page 53)

Likelihood Ratio Tests

The theory of likelihood-ratio tests is fully developed in the i.i.d. case. However, the literature on likelihood-ratio tests in the context of linear mixed models is much less extensive. First paper address the likelihood-ratio tests in linear mixed models was from Hartley and Rao (1967)

- Let $\psi = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$, $\boldsymbol{\theta}$ be the vector of variance components, $\boldsymbol{\beta}$ be the vector of fixed parameters

- A general interest is to test a subvector of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(1)}$, is identical to a known vector $\boldsymbol{\theta}_0^{(1)}$ or not

- Let $\boldsymbol{\theta}^{(2)}$ be the complement of $\boldsymbol{\theta}^{(1)}$. $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$

- Let $\hat{\boldsymbol{\theta}}$ be the global maximizer of $L(\boldsymbol{\theta}|\mathbf{y})$ over $\boldsymbol{\theta} \in \Theta$

- $\hat{\boldsymbol{\theta}}^{(2)}$ be the global maximizer of $L(\boldsymbol{\theta}_0^{(1)}, \boldsymbol{\theta}^{(2)})$ over $\boldsymbol{\theta}^{(2)} \in \Theta^{(2)}$

- $R = \dfrac{L(\boldsymbol{\theta}_0^{(1)}, \hat{\boldsymbol{\theta}}^{(2)})}{L(\hat{\boldsymbol{\theta}})}$, Hartley and Rao stated without giving a proof that $-2\log R$ is a central χ^2 with degrees of freedom r, where r is the dimension of $\boldsymbol{\theta}^{(1)}$ (Jiang 2005 proved this)

- Example discussion (Lu & Zhang 2009)

Applications

Consequences of assuming random:

- Statistical inferences can be made to the population from which the group effects were drawn
- Random effects induce a correlation among observations with the same group effect
- Estimation methods are different for fixed versus random effects
- using random effects involves making extra assumptions but often results in more precise estimates

Interpretation

- Fixed effects are interpreted as usual
- Estimates of the random effects α_i, within group, are interpreted similarly
- larger estimates of σ^2 means importance of the random effect

REML estimates v.s. ML estimates

- The default parameter estimation criterion for linear mixed models is restricted (or residual) maximum likelihood (REML)
- Maximum likelihood (ML) estimates (sometimes called “full maximum likelihood”) can be requested by specifying REML = FALSE or method=“ML” in the model
- Generally REML estimates of variance components are preferred.
- REML estimates are not guaranteed to be unbiased, but they are usually less biased than ML estimates