PANEL DATA ANALYSIS: LINEAR MIXED-EFFECT MODELS

datascience@berkeley

Theory

- For hierarchical data with a single level of grouping, we can formulate the classical LMM at a given level of a grouping factor as follows

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_\mathbf{i}$$

where $\mathbf{y_i}, \mathbf{X_i}, \beta$, and are the vectors of continuous responses, the design matrix, and the vector of residual errors for group $i, i = 1, 2, \ldots, n$, and $\mathbf{Z_i}$ and $\mathbf{b_i}$ are the matrix of covariates and the corresponding vector of random effects:

$$\mathbf{Z_i} \equiv \begin{pmatrix} \boxed{\begin{matrix} z_{i1}^{(1)} \\ \vdots \\ z_{in_i}^{(1)} \end{matrix}} & \begin{matrix} z_{i1}^{(2)} \\ \vdots \\ z_{in_i}^{(2)} \end{matrix} & \begin{matrix} \cdots \\ \ddots \\ \cdots \end{matrix} & \begin{matrix} z_{i1}^{(q)} \\ \vdots \\ z_{in_i}^{(q)} \end{matrix} \end{pmatrix}$$

$Z_i^{(k)}$

where each of the columns in $\mathbf{Z_i}$ can be more compactly represented as $\underline{\left(\mathbf{z}^{(}\mathbf{k})_\mathbf{i}\right)}$ and $\mathbf{b_i} \equiv (b_{i1}, \ldots, b_{iq})^T$

- Similar to the design matrix $\mathbf{X_i}$, the matrix $\mathbf{Z_i}$ contains known values of q covariates, with corresponding unobservable effects $\mathbf{b_i}$

Importantly, $\mathbf{b_i} \sim N_q(\mathbf{0}, \mathbf{D})$, $\epsilon_\mathbf{i} \sim N_{n_i}(\mathbf{0}, \mathbf{R_i}) with \mathbf{b_i} \perp \epsilon_\mathbf{i}$

- That is, the residual errors $\epsilon_\mathbf{i}$ for the same group are independent of the random effects $\mathbf{b_i}$.

- Also assume that vectors of random effects and residual errors for different groups are independent of each other: $b_i \perp \epsilon_{i'}$ for $i \neq i'$

$$\mathcal{D} = \sigma^2 \boldsymbol{D} \quad \text{and} \quad \mathcal{R}_i = \sigma^2 \boldsymbol{R}_i;$$

where σ^2 s an unknown scale parameter. In general, we will assume that $\mathbf{D}$ and $\mathbf{R_i}$ are positive-definite, unless stated otherwise

- In addition to the fixed-effects parameters $\mathbf{b}$ for the covariates used in constructing the design matrix $\mathbf{X_i}$, the proposed model includes two random components: the within-group residual errors $ and the random effects $\mathbf{b_i}$ for the covariates included in the matrix $\mathbf{Z_i}$. The presence of fixed and random effects of known variables gives rise to the name of this class of models.

- Some authors, researchers, partitioners refer to the proposed framework as a *two-level* or *two-stage* model. Following authors such as Pinheiro and Bates (2000), we will refer it to as a single-level linear mixed model.

- The framework established so far can easily be extended to multilevel grouped data.

- For instance, a model for data with two levels of grouping, with observations grouped into N first-level groups (indexed by $i = 1, \ldots, N$), each with n_i second-level (sub-)groups (indexed by $j = 1, \ldots, n_i$) containing n_{ij} observations, can be written as

$$\mathbf{y}_{ij} = X_{ij}\boldsymbol{\beta} + \mathbf{Z}_{1,ij}\mathbf{b}_i + \mathbf{Z}_{2,ij}\mathbf{b}_{ij} + \boldsymbol{\varepsilon}_{ij}$$

$$\mathbf{b}_i \sim \mathcal{N}_{q_1}(\mathbf{0}, \mathcal{D}_1), \quad \mathbf{b}_{ij} \sim \mathcal{N}_{q_2}(\mathbf{0}, \mathcal{D}_2), \quad \text{and} \quad \boldsymbol{\varepsilon}_{ij} \sim \mathcal{N}_{n_{ij}}(\mathbf{0}, \mathcal{R}_{ij})$$

where the random vectors $\mathbf{b}_i$, $\mathbf{b}_{ij}$, and $\boldsymbol{\varepsilon}_{ij}$ are independent of each other.

- In this setup, $\mathbf{b_i}$ are the random effects associated with the first-level groups, while $\mathbf{b_{ij}}$ are the random effects, independent of the first-level random effects, associated with the second-level groups. The design matrices $\mathbf{Z_{1,ij}}$ and $\mathbf{Z_{2,ij}}$ could be identical. Following *Pinheiro and Bates (2000)*, we refer this model as a *two-level LMM*.

- It follows that, for the classical LMMs, the conditional distribution, $f_{y|b}(\mathbf{y_i}|\mathbf{b_i})$, of $\mathbf{y_i}$ given $\mathbf{b_i}$ is multivariate normal, with the mean and variance defined as:

$$\mathrm{E}(\mathbf{y}_i|\mathbf{b}_i) \equiv \boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$$

$$\mathrm{Var}(\mathbf{y}_i|\mathbf{b}_i) = \sigma^2\boldsymbol{R}_i,$$

with $\boldsymbol{\mu}_i \equiv (\mu_{i1}, \ldots, \mu_{i,n_i})'$

$$\mathrm{E}(y_{ij}|\mathbf{b}_i) \equiv \mu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i$$

where $\mathbf{x}_{ij} \equiv (x_{ij}^{(1)}, \ldots, x_{ij}^{(p)})'$ and $\mathbf{z}_{ij} \equiv (z_{ij}^{(1)}, \ldots, z_{ij}^{(q)})'$ are column vectors

- Conditionally on the (unknown) values of the random effects $\mathbf{b_i}$, the mean value of the dependent-variable vector $\mathbf{y_i}$ is defined by a linear combination of the vectors of the $\mathbf{X}$- and $\mathbf{Z}$-covariates included, as columns, in the group-specific design matrices $\mathbf{X_i}$ and $\mathbf{Z_i}$, corresponding to the fixed effects $ and random effects $\mathbf{b_i}$.

- Moreover, the conditional variance-covariance matrix of $\mathbf{y_i}$ is equal to the variance-covariance matrix of the residual errors $\epsilon_\mathbf{i}$

Maximum-Likelihood Estimation

$$\ell_{\text{Full}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) \equiv -\frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\log[\det(\boldsymbol{V}_i)]$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\mathbf{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) \equiv \left(\sum_{i=1}^{N}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{N}\boldsymbol{X}_i'\boldsymbol{V}_i^{-1}\mathbf{y}_i,$$

$$\widehat{\sigma}^2_{\text{ML}}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N}\mathbf{r}_i'\boldsymbol{V}_i^{-1}\mathbf{r}_i/n,$$

where $\mathbf{r}_i \equiv \mathbf{r}_i(\boldsymbol{\theta}) = \underbrace{\mathbf{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})}$.

$$\widehat{\sigma}^2_{\text{REML}}(\boldsymbol{\theta}) \equiv \sum_{i=1}^{N} \mathbf{r}_i' V_i^{-1} \mathbf{r}_i / (n-p)$$

- This leads to a log-profile-restricted-likelihood function, which only depends on θ:

$$\ell^*_{\text{REML}}(\boldsymbol{\theta}) \equiv -\frac{n-p}{2} \log\left(\sum_{i=1}^{N} \mathbf{r}_i' \mathbf{r}_i\right) - \frac{1}{2} \sum_{i=1}^{N} \log[\det(\boldsymbol{V}_i)]$$

$$-\frac{1}{2} \log\left[\det\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)\right].$$

To sum up, we have:

$$\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{o}, \sigma^2 \boldsymbol{W}^{-1})$$

$$(\mathcal{Y}|\mathcal{B} = \boldsymbol{b}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{o}, \sigma^2 \boldsymbol{W}^{-1})$$

$\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ a parameterized $q \times q$ variance-covariance matrix, $\boldsymbol{\Sigma}$;

As a variance-covariance matrix, $\boldsymbol{\Sigma}$ must be positive semidefinite.

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \sigma^2 \boldsymbol{\Lambda}_{\boldsymbol{\theta}} \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^\top;$$

a *relative covariance factor*, $\boldsymbol{\Lambda}_{\boldsymbol{\theta}}$, which is a $q \times q$ matrix;

Final Remarks:

- This formulation of linear mixed models allows for a relatively compact expression for the profiled log-likelihood of θ

- The matrices associated with random effects, $\mathbf{Z}$ and $\mathbf{\Lambda}_\theta$, typically have a sparse structure with a sparsity pattern that encodes various model assumptions.

- More on the details on the structure and how to represent it efficiently can be found in the assigned readings, R demo, or the live sessions.

- The setup above is very general. The interface provided by *lme4*'s *lmer* function is less general than the model described.

- To take advantage of the entire range of possibilities, one may have to use the modular functions or explore the experimental *flexLambda* branch of *lme4* on *Github*

Berkeley SCHOOL OF INFORMATION