Lecture 9

Mixed models

Mixed models extend linear regression models in that additionally to the fixed coefficients (=fixed effects) also random effects are included (hence, the name: mixed effects models). Mixed models are used to model clustered, hierarchical or longitudinal data, where the random effects typically model some unobserved source of heterogeneity. For more details on these models consult Demidenko, E. (2013) *Mixed models: Theory and applications* and Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications*.

The data we consider are of the form Y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, k_i$. That is, there are n subjects (clusters) and for each subject i there are k_i observations. Longitudinal data are observations of n subjects which are taken over time. A typical example of a longitudinal study is a measurement of some quantity of interest (hight, weight, blood pressure, etc.) over time on n subjects. In clustered data the observations are typically taken at one time point, but the observations can be grouped by some feature.

A linear mixed model is written as

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i, \ i = 1, \dots, n,$$

where $Y_i, \epsilon_i \in \mathbb{R}^{k_i}$, $X_i \in \mathbb{R}^{k_i \times p}$, $Z_i \in \mathbb{R}^{k_i \times m}$, $\beta \in \mathbb{R}^p$ are fixed coefficients (effects) and $u_i \in \mathbb{R}^m$ are random variables with zero mean, which are independent on ϵ_i. That is, all subjects have the same fixed effect β and different random effects u_i. Fixed effects β are also called marginal or population average effects. The linear mixed model can also be represented as

$$Y = X\beta + Zu + \epsilon, \ \ \mathrm{E}\begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \ \mathrm{cov}\begin{pmatrix} u \\ \epsilon \end{pmatrix} = \begin{pmatrix} \sigma_u^2 D & 0 \\ 0 & \sigma_\epsilon^2 \Sigma \end{pmatrix},$$

for $Y, \epsilon \in \mathbb{R}^N$, $N = \sum_{i=1}^n k_i$, $X \in \mathbb{R}^{N \times p}$, $Z = \mathrm{blockdiag}(Z_1, \dots, Z_n) \in \mathbb{R}^{N \times nm}$, $u \in \mathbb{R}^{nm}$. Typically, a normality assumption is made both for ϵ and u.

Examples

1. One-way ANOVA (analysis of variance) model is typically written as

$$Y_{ij} = \beta_i + \epsilon_{ij}, \ \ i = 1, \dots, n, \ \ j = 1, \dots, k_i.$$

for fixed β_i, which sometimes also represented as $\beta_i = \beta + u_i$, where both β and u_i are fixed. Assuming that u_i are random, for example, normally distributed, leads to a model also known as variance components model, which is a mixed model:

$$Y_{ij} = \beta + u_i + \epsilon_{ij}, \;\; \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \;\; u_i \sim \mathcal{N}(0, \sigma_u^2), \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, k_i.$$

We can represent this model in the matrix notation as above setting $X = 1_N$, $Z = \text{blockdiag}(1_{k_1}, \ldots, 1_{k_n})$.

2. Growth curve models typically have the form

$$Y_{ij} = X_i(\beta + u_i) + \epsilon_{ij}, \;\; \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \;\; u_i \sim \mathcal{N}_p(0_p, \sigma_u^2 I_p), \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, k_i,$$

where $X_i \in \mathbb{R}^{n \times p}$ represents the growth pattern. For example, if Y_{ij} is a height of an ith child at time point t_j, then one could take $X_i = (1, t_i, t_i^2)$.

3. In small area estimation is assumed that the means in "small areas" differ for a random amount:

$$Y_i = X_i\beta + 1_{n_i} u_i + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}_{k_i}(0_{k_i}, \sigma^2 I_{k_i}), \;\; u_i \sim \mathcal{N}(0, \sigma_u^2), \;\; i = 1, \ldots, n,$$

where in each ith small area there are k_i observations available.

Estimation in mixed models can be carried out using the maximum likelihood approach, if distributional assumptions on ϵ and u are made. To make the model identifiable we assume that $\sum_{i=1}^n X_i^t X_i$ is non-singular and that $N = \sum_{i=1}^n k_i > p$. Moreover, at least one matrix $Z_i^t Z_i$ is positive definite and $\sum_{i=1}^n (k_i - m) = N - nm > 0$.

Under normality assumptions $\epsilon \sim \mathcal{N}_N(0_N, \sigma_\epsilon^2 \Sigma)$ and $u_i \sim \mathcal{N}_m(0_m, \sigma_u^2 \widetilde{D})$ we can represent the mixed model as

$$Y \sim \mathcal{N}_N(X\beta, \sigma_\epsilon^2 \Sigma + \sigma_u^2 Z D Z^t),$$

where $D = \text{blockdiag}(\widetilde{D}, \ldots, \widetilde{D}) \in \mathbb{R}^{nm \times nm}$, so that the log-likelihood for the parameter vector $\theta = (\beta, u, \sigma_\epsilon^2, \sigma_u^2, \Sigma, D)$ becomes

$$\ell(\theta) = -\frac{1}{2} \left\{ N \log(2\pi\sigma_\epsilon^2) + \log|V| + \frac{(Y - X\beta)^t V^{-1} (Y - X\beta)}{\sigma_\epsilon^2} \right\},$$

where $V = \Sigma + \sigma_u^2 Z D Z^t / \sigma_\epsilon^2$. It is often assumed that $\Sigma = I_N$. Assumption that $\widetilde{D} = I_m$ is not justified if the observations for each subject (=within a cluster) are taken over time, but is reasonable in many other settings. If $\widetilde{D} \neq I_m$, then some parametric covariance structure is assumed (e.g. an autoregressive process), so that estimation of $\widetilde{D}$ is reduced to estimation of several parameters. In the following we consider a simple case $\widetilde{D} = I_m$

and $\Sigma = I_N$, so that $V = V(\lambda) = I_N + ZZ^t/\lambda$ for $\lambda = \sigma_\epsilon^2/\sigma_u^2$.

It is easy to see that the estimator for β is given by

$$\widehat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} Y,$$

which is a generalised least squares estimator. Note that this estimator depends on the unknown λ. Estimator for σ_ϵ^2 results in

$$\widehat{\sigma}_\epsilon^2 = \frac{(Y - X\widehat{\beta})^t V^{-1} (Y - X\widehat{\beta})}{N} = \frac{Y^t V^{-1} (Y - X\widehat{\beta})}{N}.$$

It remains to estimate λ. Using

$$\begin{aligned} \frac{\partial \log(|V|)}{\partial \lambda} &= \operatorname{tr}\left(V^{-1} \frac{\partial V}{\partial \lambda} \right), \quad \frac{\partial V}{\partial \lambda} = -\frac{1}{\lambda^2} ZZ^t \\ \frac{\partial V^{-1}}{\partial \lambda} &= -V^{-1} \frac{\partial V}{\partial \lambda} V^{-1} \\ V^{-1} &= I_N - Z(Z^t Z + \lambda I_{nm})^{-1} Z^t \end{aligned}$$

we get the estimating equation from the profile likelihood

$$\begin{aligned} \frac{-2\partial \ell(\lambda; \widehat{\beta}, \widehat{\sigma}_\epsilon^2)}{\partial \lambda} &= \frac{1}{\widehat{\sigma}_\epsilon^2} \frac{\partial \widehat{\sigma}_\epsilon^2}{\partial \lambda} - \frac{1}{\lambda^2} \operatorname{tr}\left[\{ I_N - Z(Z^t Z + \lambda I_{nm})^{-1} Z^t \} ZZ^t \right] \\ &= \frac{1}{\lambda \widehat{\sigma}_\epsilon^2} Y^t S(I_N - S) Y - \frac{1}{\lambda} \operatorname{tr}\{ (Z^t Z + \lambda I_{nm})^{-1} Z^t Z \}, \end{aligned}$$

where $S = C(C^t C + \lambda J)^{-1} C^t$ for $C = (X, Z)$ and $J = \operatorname{diag}(0_p, 1_{nm})$. Finding a solution to

$$\frac{\partial \ell(\lambda; \widehat{\beta}, \widehat{\sigma}_\epsilon^2)}{\partial \lambda} = 0$$

gives an estimator for λ and hence σ_ϵ^2 and β.

It is easy to see that the maximum likelihood estimators for σ_ϵ^2 is biased. This in turn leads to inferior estimators of λ and β. It has been suggested in Patterson and Thompson (1971) and later in Harville (1974) to use the so-called *restricted* (also called *residual*) maximum likelihood. This approach is best justified in the Bayesian framework putting a flat non-informative prior on β. The restricted likelihood is defined to be

$$\ell_R(\theta) = \ell(\theta) - \frac{1}{2} \log |\sigma_\epsilon^2 X^t V^{-1} X|.$$

The restricted maximum likelihood (REML) estimator for β (with known λ) is unchanged, the estimator for σ_ϵ^2 becomes

$$\widehat{\sigma}_\epsilon^2 = \frac{Y^t V^{-1} (Y - X\widehat{\beta})}{N - p}$$

and λ is found by solving

$$\frac{-2\partial\ell(\lambda;\widehat{\beta},\widehat{\sigma}^2_\epsilon)}{\partial\lambda} = \frac{1}{\lambda\widehat{\sigma}^2_\epsilon}Y^tS(I_N-S)Y - \frac{1}{\lambda}\{\mathrm{tr}(S)-p\} = 0.$$

Once λ is obtained, $\widehat{\beta}$ follows immediately and finally an estimator for $\widehat{\sigma}^2_\epsilon$ is calculated. Conditions for the existence of (restricted) maximum likelihood estimators in linear mixed models are given Demidenko (2013). Under mild regularity conditions all asymptotic properties of maximum likelihood estimators (consistency, asymptotic normality) apply.

We have estimated fixed parameters in the model and herewith the marginal effect β. However, in practice the conditional effects are also of interest. Since u is assumed to be random, one typically uses the term "predictor" for $\widehat{u}$, which is obtained as a *best linear unnbiased predictor* (BLUP). Henderson (1950) suggested to maximise the joint density (up to a constant) of u and Y in order to get estimators for β and u:

$$\left|\begin{pmatrix}\sigma^2_u D & 0\\ 0 & \sigma^2_\epsilon\Sigma\end{pmatrix}\right|^{-1/2}\exp\left\{-\frac{1}{2}\begin{pmatrix}u\\ \epsilon\end{pmatrix}^t\begin{pmatrix}\sigma^2_u D & 0\\ 0 & \sigma^2_\epsilon\Sigma\end{pmatrix}^{-1}\begin{pmatrix}u\\ \epsilon\end{pmatrix}\right\},$$

where $\epsilon = Y - X\beta - Zu$. This leads to the criterion (up to terms independent of β and u)

$$(Y-X\beta-Zu)^t(\sigma^2_\epsilon\Sigma)^{-1}(Y-X\beta-Zu) + u^t(\sigma^2_u D)^{-1}u.$$

Note that this expression highlights that u are penalised. Setting as before $\Sigma = I_N$, $D = I_{nm}$ and $\lambda = \sigma^2_\epsilon/\sigma^2_u$ we get the criterion

$$(Y-X\beta-Zu)^t(Y-X\beta-Zu) + \lambda u^t u = (Y-C\alpha)^t(Y-C\alpha) + \lambda\alpha^t J\alpha,$$

for $\alpha = (\beta, u)$. From this criterion we obtain for a known λ

$$\begin{aligned}\widehat{\alpha} &= (C^tC+\lambda J)^{-1}C^tY\\ \widehat{\beta} &= (X^tV^{-1}X)^{-1}X^tY\\ \widehat{u} &= \frac{1}{\lambda}Z^tV^{-1}(Y-X\widehat{\beta}).\end{aligned}$$

The predictor $\widehat{u}$ is also the best linear unbiased predictor of u (for known β and λ), which can be seen as follows. Assume we would like to obtain the best linear predictor of u from $Y - X\beta = Zu + \epsilon$. That is, we aim to find

$$\arg\min_{a,B}\mathrm{E}\|Zu - \{a + B(Y-X\beta)\}\|^2,$$

which are

$$\begin{aligned}\widehat{a} &= \mathrm{E}(Zu) - E\{B(Y-X\beta)\} = 0\\ \widehat{B} &= \mathrm{cov}(Zu, Y-X\beta)\mathrm{cov}(Y-X\beta)^{-1} = Z\mathrm{cov}(u)Z^t\mathrm{cov}(Y)^{-1} = \sigma^2_u ZZ^tV^{-1}/\sigma^2_\epsilon,\end{aligned}$$

leading to

$$\widehat{u} = \frac{1}{\lambda} Z^t V^{-1}(Y - X\beta).$$

Once β and λ are replaced by estimators, $\widehat{u}$ is, strictly speaking, not BLUP.

Inference in mixed models is based on asymptotic normality.