Linear Mixed Effects Models
Inference

Biostatistics 653

Applied Statistics III: Longitudinal Data Analysis

Estimation via Maximum Likelihood

We consider a standard linear mixed effects model

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

where b_i is independent of ϵ_i, with $b_i \sim MVN(0, D)$ and $\epsilon_i \sim MVN_{n_i}(0, R_i)$.

For the sake of simplicity, let $R_i = \sigma^2 I_{n_i}$. Then

$$L(\beta, D, \sigma^2) \propto \prod_{i=1}^{N} |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(Y_i - X_i\beta)^T \Sigma_i^{-1}(Y_i - X_i\beta)}$$

where $\Sigma_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$

Estimation via Maximum Likelihood

- The standard approach is to take the 1st and 2nd derivatives of the loglikelihood and (a) use Newton-Raphson (based on observed information) or (b) Fisher scoring (based on expected information). (c) Often a hybrid estimation approach is used, because conditional on knowing Σ_i, then

$$\hat{\beta}_{ML} = \left(\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^T \Sigma_i^{-1} Y_i\right)$$

- So given $\left(\widehat{D}, \hat{\sigma}^2\right)^{(t)}$ estimated from the t-th iteration, we get $\hat{\beta}^{(t)}$ and use it to calculate $\left(\widehat{D}, \hat{\sigma}^2\right)^{(t+1)}$, iterating until convergence.

Introduction to the EM Algorithm

- We consider a general optimization problem, where we want to maximize the (marginal) likelihood $L(\theta; X) = P(X|\theta)$.
- Our observed data is X and parameter is θ. To facilitate computation, we introduce a set of missing data Z.
- The complete data likelihood is

$$L(\theta; X, Z) = P(X, Z|\theta)$$

- And the marginal likelihood can be expressed as

$$L(\theta; X) = P(X|\theta) = \sum_{Z} P(X, Z|\theta)$$

Introduction to the EM Algorithm

- The goal of the EM algorithm is to find MLE estimates for the marginal likelihood via optimization iterations using the complete likelihood.
- The EM algorithm consists of two steps:
- The Expectation Step (E-Step)

$$Q\left(\theta \middle| \theta^{(t)}\right) = E_{Z|X,\theta^{(t)}}(\log L(\theta; X, Z))$$

- The Maximization Step (M-Step)

$$\theta^{(t+1)} = argmax\_\theta \; Q\left(\theta \middle| \theta^{(t)}\right)$$

Introduction to the EM Algorithm

- EM works to improve $Q\left(\theta \middle| \theta^{(t)}\right)$, which implies improvement to the log likelihood $\log L(\theta; X)$.
- To see this, we have

$$\log P(X|\theta) = \log P(X, Z|\theta) - \log P(Z|X, \theta)$$
$$= (\textstyle\sum_Z P(Z|X, \theta^{(t)}))(\log P(X, Z|\theta) - \log P(Z|X, \theta))$$
$$= Q\left(\theta \middle| \theta^{(t)}\right) + H(\theta|\theta^{(t)})$$

where $Q\left(\theta \middle| \theta^{(t)}\right)$ is the expected value of $\log P(X, Z|\theta)$ with respect to the conditional distribution $P(Z|X, \theta^{(t)})$; while $H(\theta|\theta^{(t)})$ is the expected value of $-\log P(Z|X, \theta)$ with respect to the conditional distribution $P\left(Z \middle| X, \theta^{(t)}\right)$.

Introduction to the EM Algorithm

- Replacing θ with $\theta^{(t)}$ we have

$$\log P(X|\theta^{(t)}) = Q\left(\theta^{(t)}\middle|\theta^{(t)}\right) + H(\theta^{(t)}|\theta^{(t)})$$

- Therefore

$$\log P(X|\theta) - \log P(X|\theta^{(t)})$$
$$= Q\left(\theta\middle|\theta^{(t)}\right) - Q\left(\theta^{(t)}\middle|\theta^{(t)}\right) + H\left(\theta\middle|\theta^{(t)}\right) - H(\theta^{(t)}|\theta^{(t)})$$

- Based on Gibbs inequality, we have

$$H\left(\theta\middle|\theta^{(t)}\right) = -\sum_{Z} P\left(Z\middle|X,\theta^{(t)}\right) \log P(Z|X,\theta)$$
$$> -\sum_{Z} P\left(Z\middle|X,\theta^{(t)}\right) \log P\left(Z\middle|X,\theta^{(t)}\right) = H(\theta^{(t)}|\theta^{(t)})$$

- Therefore,

$$\log P(X|\theta) - \log P\left(X\middle|\theta^{(t)}\right) > Q\left(\theta\middle|\theta^{(t)}\right) - Q\left(\theta^{(t)}\middle|\theta^{(t)}\right)$$

- We just need to choose a θ that increases $Q\left(\theta\middle|\theta^{(t)}\right)$ on top of $Q\left(\theta^{(t)}\middle|\theta^{(t)}\right)$, which will improve $\log P(X|\theta)$.

Estimation using the EM Algorithm

- Estimation in linear mixed effects models may also be carried out using the EM algorithm or using REML instead of ML.
- In the EM algorithm, the observed data is
$$Y = (Y_1^T, \cdots, Y_N^T)$$
- We view b_i and ϵ_i as the missing data. The complete data is therefore
$$X = (Y_1^T, b_1^T, \epsilon_1^T, \cdots, Y_N^T, b_N, \epsilon_N^T)$$
- The joint distribution of $(Y_i^T, b_i^T, \epsilon_i^T)$ is the multivariate normal:

$$\begin{pmatrix} Y_i \\ b_i \\ \epsilon_i \end{pmatrix} \sim N(\begin{pmatrix} X_i\beta \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Z_i D Z_i^T + \sigma^2 I_{n_i} & Z_i D & \sigma^2 I_{n_i} \\ D Z_i^T & D & 0 \\ \sigma^2 I_{n_i} & 0 & \sigma^2 I_{n_i} \end{pmatrix})$$

Estimation using the EM Algorithm

- The above joint distribution is based on:

$$Cov(Y_i, b_i) = Cov(X_i\beta + Z_i b_i + \epsilon_i, b_i) = Cov(Z_i b_i, b_i) = Z_i D$$
$$Cov(Y_i, \epsilon_i) = Cov(X_i\beta + Z_i b_i + \epsilon_i, \epsilon_i) = Cov(\epsilon_i, \epsilon_i) = \sigma^2 I_{n_i}$$
$$Cov(b_i, \epsilon_i) = 0$$

- Because of the constraint $Y_i = X_i\beta + Z_i b_i + \epsilon_i$, the covariance matrix in the previous slide is singular. In addition, given b_i and ϵ_i, Y_i contributes nothing to the estimation of D and σ^2.

Estimation using the EM Algorithm

- The complete data likelihood is given by

$$\prod_{i=1}^{N} P(Y_i, b_i, \epsilon_i | \beta, D, \sigma^2) = \prod_{i=1}^{N} P(Y_i | b_i, \epsilon_i, \beta) P(b_i | D) P(\epsilon_i | \sigma^2)$$

$$\propto \prod_{i=1}^{N} |D|^{-\frac{1}{2}} e^{-\frac{1}{2} b_i^T D^{-1} b_i} (\sigma^2)^{-\frac{n_i}{2}} e^{-\frac{1}{2\sigma^2} \epsilon_i^T \epsilon_i}$$

- Thus, the complete data sufficient statistics for D and σ^2 are given by $\sum_{i=1}^{N} b_i b_i^T$ and $\sum_{i=1}^{N} \epsilon_i^T \epsilon_i$.

Estimation using the EM Algorithm

- The E-Step is

$$Q = \sum_{i=1}^{N} E(-\frac{1}{2}\log|D| - \frac{1}{2} b_i^T D^{-1} b_i - \frac{n_i}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\epsilon_i^T\epsilon_i)$$

where the expectation is taken with respect to the conditional distribution $\epsilon_i, b_i \middle| Y_i, D^{(t)}, \left(\sigma^2\right)^{(t)}, \beta^{(t)}$.

- We maximize Q by setting the first two derivatives to 0:

$$\frac{\partial Q}{\partial(\sigma^2)^{-1}} = \sum_{i=1}^{N} \frac{n_i}{2}\sigma^2 - \frac{1}{2}E(\epsilon_i^T\epsilon_i) = 0$$

$$\frac{\partial Q}{\partial D^{-1}} = \sum_{i=1}^{N} \frac{1}{2}D - \frac{1}{2}E(b_i b_i^T) = 0$$

Estimation using the EM Algorithm

- The M-Step is therefore

$$\hat{\sigma}^2 = \sum_{i=1}^{N} E(\epsilon_i^T \epsilon_i) \,/ \sum_{i=1}^{N} n_i$$

$$\hat{D} = \sum_{i=1}^{N} E(b_i b_i^T) \,/N$$

- To obtain these expectations, we will use the relationship

$$E\left(b_i b_i^T \middle| Y_i, \beta, D, \sigma^2\right)$$
$$= E(b_i | Y_i, \beta, D, \sigma^2) E\left(b_i^T \middle| Y_i, \beta, D, \sigma^2\right) + V(b_i | Y_i, \beta, D, \sigma^2)$$

- Thus, we need to calculate $E(b_i|Y_i, \beta, D, \sigma^2)$ and $V(b_i|Y_i, \beta, D, \sigma^2)$.

Estimation using the EM Algorithm

- Recall that, for

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

- We know

$$X_1 | X_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$$

Where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Estimation using the EM Algorithm

- In our case, we have

$$\binom{Y_i}{b_i} \sim MVN(\binom{X_i\beta}{0}, \begin{pmatrix} Z_i D Z_i^T + \sigma^2 I_{n_i} & Z_i D \\ D Z_i^T & D \end{pmatrix})$$

- Thus

$$E(b_i|Y_i, \beta, D, \sigma^2) = DZ_i^T \Sigma_i^{-1}(Y_i - X_i\beta)$$
$$\mathrm{V}(b_i|Y_i, \beta, D, \sigma^2) = D - DZ_i^T \Sigma_i^{-1} Z_i D$$

$$E\left(b_i b_i^T \middle| Y_i, \beta, D, \sigma^2\right)$$
$$= DZ_i^T \Sigma_i^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T \Sigma_i^{-1} Z_i D + D - DZ_i^T \Sigma_i^{-1} Z_i D$$

Estimation using the EM Algorithm

- Similarly

$$E(\epsilon_i|Y_i,\beta,D,\sigma^2) = \sigma^2 I_{n_i}\Sigma_i^{-1}(Y_i - X_i\beta)$$
$$\mathrm{V}(\epsilon_i|Y_i,\beta,D,\sigma^2) = \sigma^2 I_{n_i} - \sigma^2 I_{n_i}\Sigma_i^{-1}\sigma^2 I_{n_i}$$

- Thus

$$E\left(\epsilon_i\epsilon_i^T \middle| Y_i,\beta,D,\sigma^2\right)$$
$$= \sigma^4\Sigma_i^{-1}(Y_i - X_i\beta)(Y_i - X_i\beta)^T\Sigma_i^{-1} + \sigma^2 I_{n_i} - \sigma^4\Sigma_i^{-1}$$

$$E\left(\epsilon_i^T\epsilon_i \middle| Y_i,\beta,D,\sigma^2\right) = trE\left(\epsilon_i\epsilon_i^T \middle| Y_i,\beta,D,\sigma^2\right)$$

Estimation using the EM Algorithm

- So given the observed data and the current values of the parameter estimates, we can calculate $\sum_{i=1}^{N} E(\epsilon_i^T \epsilon_i)$ and $\sum_{i=1}^{N} E(b_i b_i^T)$ for the E-step, go back to the M-step, etc. and iterate until convergence.

Prediction of Random Effects

It is often useful to estimate individual subject effects

- to plot an individual's growth curve,
- to estimate an individual's mean value, or
- to identify individuals with the largest (smallest) outcomes or who grow fastest (slowest), etc.

- If n_i were large, it might be reasonable to treat the bi as fixed effects. However, often some subjects have quite small n_i, and we generally must borrow information from other subjects in order to predict the random effects.

Prediction of Random Effects

Two approaches:

- Extend Gauss-Markov theorem to random effects (cf. Harville, 1976 Annals of Statistics, 1977 JASA)
- Empirical Bayes approach (cf. Laird and Ware, 1982 Biometrics)

These are equivalent in linear models under multivariate normality.

Prediction of Random Effects

Recall that for the univariate linear model,

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I)$$

the best linear unbiased estimator (BLUE) $C\hat{\beta}$ of any estimable contrast $C\beta$ satisfies

- $E\left(C\hat{\beta}\right) = C\beta$ and
- $V\left(C\hat{\beta}\right) \leq Var(C\hat{\beta}^*)$ for any other linear unbiased estimator $\hat{\beta}^*$

Prediction of Random Effects

- To extend this to random effects, let $b^T = (b_1^T, \cdots, b_N^T)$ denote the vector of all the random effects. Then the best linear unbiased predictor (BLUP) of $C_1\beta + C_2 b$ is the linear function of Y that is unbiased and has minimum variance in the class of unbiased linear predictors.
- In this case, it can be shown that (see Henderson, Harville, Searle, and others) we can predict b_i using its conditional mean as
$$E\left(b_i \middle| Y_i, \hat{\beta}, D, \sigma^2\right) = D Z_i^T \Sigma_i^{-1}(Y_i - X_i\hat{\beta})$$
- Thus the BLUP estimator depends on D and R_i, usually estimated by ML or REML. So plugging in $\widehat{D}$ and $\hat{R}_i$, estimates of these variance components, gives us an estimator close to the BLUP of bi, which is often called the “empirical BLUP”.

Prediction of Random Effects

- The variance of the empirical BLUP is usually estimated by

$$\hat{V}(\hat{b}_i)$$
$$= \hat{D} - \hat{D} Z_i^T \hat{\Sigma}_i^{-1} Z_i \hat{D} + \hat{D} Z_i^T \hat{\Sigma}_i^{-1} X_i \left(\sum_{i=1}^{N} X_i^T \, \hat{\Sigma}_i^{-1} X_i \right)^{-1} X_i \hat{\Sigma}_i^{-1} Z_i \hat{D}$$

by using the formular $V(A) = E\big(V(A|B)\big) + V(E(A|B))$

- We calculate the ith subject's predicted response profile (BLUP of the individual specific mean) as

$$\hat{Y}_i = X_i \hat{\beta} + Z_i \hat{b}_i = X_i \hat{\beta} + Z_i \hat{D} Z_i^T \hat{\Sigma}_i^{-1} (Y_i - X_i \hat{\beta})$$

Prediction of Random Effects

- Now, let us consider how the BLUP is a "shrinkage" estimator. Starting with our EB predictor
$$\hat{b}_i = DZ_i^T \Sigma_i^{-1}(Y_i - X_i\hat{\beta})$$
- We use the following identities based on Woodbury formula. Assuming $\Sigma_i = \sigma^2 I + Z_i D Z_i^T$, we have

$$\Sigma_i^{-1} = \frac{1}{\sigma^2}(I - \frac{1}{\sigma^2} Z_i \left(D^{-1} + \frac{Z_i^T Z_i}{\sigma^2}\right)^{-1} Z_i^T)$$
$$Z_i^T \Sigma_i^{-1} = \frac{1}{\sigma^2} D^{-1} \left(D^{-1} + \frac{Z_i^T Z_i}{\sigma^2}\right)^{-1} Z_i^T$$

Prediction of Random Effects

- Using these identities, we write our EB predictor as

$$\hat{b}_i = \frac{D}{\sigma^2}\left(I + \frac{Z_i^T Z_i D}{\sigma^2}\right)^{-1} Z_i^T\left(Y_i - X_i\hat{\beta}\right)$$

$$= \frac{D}{\sigma^2}\left(I + \frac{Z_i^T Z_i D}{\sigma^2}\right)^{-1} \left(Z_i^T Z_i\right)\left(Z_i^T Z_i\right)^{-1} Z_i^T\left(Y_i - X_i\hat{\beta}\right)$$

- And we note that $\left(Z_i^T Z_i\right)^{-1} Z_i^T\left(Y_i - X_i\hat{\beta}\right)$ is just a least squares estimator of b_i, obtained by taking the individual residuals $R_i = Y_i - X_i\hat{\beta}$, so we will call this b_i^{LS}.
- Then, we write

$$\hat{b}_i = D\left(I + \frac{Z_i^T Z_i D}{\sigma^2}\right)^{-1} \frac{\left(Z_i^T Z_i\right)}{\sigma^2} b_i^{LS} = D\left(\sigma^2\left(Z_i^T Z_i\right)^{-1} + D\right)^{-1} b_i^{LS}$$

$$= A b_i^{LS}$$

Prediction of Random Effects

- Then, we write

$$\hat{b}_i = D\left(I + \frac{Z_i^T Z_i D}{\sigma^2}\right)^{-1} \frac{\left(Z_i^T Z_i\right)}{\sigma^2} b_i^{LS}$$
$$= D\left(\sigma^2\left(Z_i^T Z_i\right)^{-1} + D\right)^{-1} b_i^{LS} = A b_i^{LS}$$

- where A = (prior variance)(total variance)$^{-1}$ with D as the prior variance and $\left(\sigma^2\left(Z_i^T Z_i\right)^{-1} + D\right)$ equal to the total variance (prior variance plus the least squares variance $\sigma^2\left(Z_i^T Z_i\right)^{-1}$.

Prediction of Random Effects

- Because A is only a fraction of the total variance, $\hat{b}_i$ is "shrunken" relative to b_i^{LS}.
- If we have lots of information in the data Y_i, then $\sigma^2\left(Z_i^T Z_i\right)^{-1} \rightarrow 0$, so that $A \rightarrow I$ and $\hat{b}_i \rightarrow b_i^{LS}$.
- When the information in the data is poor relative to the prior, then $\hat{b}_i \approx 0$, and our individual predictions $\hat{Y}_i = X_i\hat{\beta} + Z_i\hat{b}_i$ rely more on population average estimates than individual-specific estimates.

Prediction of Random Effects

- The shrinkage estimates of $\hat{Y}_i$ are a compromise between population average estimates $X_i\hat{\beta}$ and individual-specific OLS estimates from separate models, $X_i\hat{\beta}_i$, with the degree of reliance on the population or individual-specific estimates depending on how much variability in the data is within versus between subjects.

Bayesian Motivation

- A Bayesian approach treats

$$b_i \sim N_q(0, D)$$

as a prior for b_i. Given the data Y_i, we compute the conditional posterior of b_i (the conditional distribution of b given the observed data and other parameters), treating the covariance parameters as fixed and using a flat prior for β.

- The empirical Bayes approach uses observed data to estimate D (rather than specifying a prior distribution for D). While the Bayesian approach provides us with the entire posterior distribution of b_i (and not just a point estimate), we can show the estimate of b_i taken as the mean of its estimated conditional posterior distribution is the same as that given previously.
- In practice, we have replication coming from Y_i and thus can estimate β, R_i and D from the data. Because we replace the unknown parameters by their maximum or restricted maximum likelihood estimates, we obtain "empirical" Bayes estimates.