Fitting a Linear Mixed-Effects Model using Expectation-Maximization

Camden Lopez

2022-06-15

Problem Statement

Assume that for $i = 1, \ldots, n$,

- $Y_i = X_i\beta + Z_i b_i + \epsilon_i$
- X_i is a fixed $n_i \times p$ matrix
- Z_i is a fixed $n_i \times q$ matrix
- $b_i \sim \mathrm{N}_q(0, \Sigma_b)$,
- $\epsilon_i \sim \mathrm{N}_{n_i}(0, \sigma^2 I_{n_i})$
- b_i and ϵ_i are independent
- (b_i, ϵ_i) and (b_j, ϵ_j) are independent for all $i \neq j$

Given (X_i, Z_i, Y_i) for $i = 1, \ldots, n$, we want to estimate β, Σ_b, and σ^2 using the expectation-maximization (EM) algorithm.

The likelihood function is

$$L(\beta, \Sigma_b, \sigma^2) = \prod_{i=1}^{n} f(y_i|b_i, \beta, \sigma^2) g(b_i|\Sigma_b)$$

where f is the density of y_i conditional on b_i, and g is the density of b_i.

$$\begin{aligned}
f(y_i|b_i, \beta, \sigma^2) &= (2\pi)^{-n_i/2}(\sigma^2)^{-n_i/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - X_i\beta - Z_i b_i)^\intercal (y_i - X_i\beta - Z_i b_i)\right] \\
g(b_i|\Sigma_b) &= (2\pi)^{-q/2}|\Sigma_b|^{-1/2} \exp\left[-\frac{1}{2} b_i^\intercal \Sigma_b^{-1} b_i\right]
\end{aligned}$$

The log-likelihood is

$$
\begin{aligned}
l(\beta, \Sigma_b, \sigma^2) &= \sum_{i=1}^{n} \Big[\log f(y_i|b_i, \beta, \sigma^2) + \log g(b_i|\Sigma_b)\Big] \\
&= \sum_{i=1}^{n}[-\frac{n_i}{2}\log(2\pi) - \frac{n_i}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y_i - X_i\beta - Z_ib_i)^\intercal(y_i - X_i\beta - Z_ib_i) \\
&\quad - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_b| - \frac{1}{2}b_i^\intercal\Sigma_b^{-1}b_i]
\end{aligned}
$$

E-Step Derivation

For the EM algorithm, first we set some initial values of the parameters, $\beta^{(0)}$, $\Sigma_b^{(0)}$, and $\sigma^{2(0)}$. Then, given current estimates $\beta^{(k)}$, $\Sigma_b^{(k)}$, and $\sigma^{2(k)}$, we calculate $Q(\beta, \Sigma_b, \sigma^2|\beta^{(k)}, \Sigma_b^{(k)}, \sigma^{2(k)})$, the expectation of $l(\beta, \Sigma_b, \sigma^2)$ over the distribution of $b = (b_1, \ldots, b_n)$ conditional on $y = (y_1, \ldots, y_n)$.

The conditional distribution of $b_i|y_i$ has pdf $f(y_i|b_i)g(b_i)/h(y_i)$ where h is the marginal pdf of y_i. We find the distribution of $b_i|y_i$ by looking at the kernel of $f(y_i|b_i)g(b_i)/h(y_i)$:

$$
\begin{aligned}
\frac{f(y_i|b_i)g(b_i)}{h(y_i)} &\propto \exp\left\{-\frac{1}{2\sigma^2}(y_i - X_i\beta - Z_ib_i)^\intercal(y_i - X_i\beta - Z_ib_i) - \frac{1}{2}b_i^\intercal\Sigma_b^{-1}b_i\right\} \\
&\text{(letting } r_i = y_i - X_i\beta) \\
&= \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}r_i^\intercal r_i - \frac{2}{\sigma^2}b_i^\intercal Z_i^\intercal r_i + \frac{1}{\sigma^2}b_i^\intercal Z_i^\intercal Z_ib_i + b_i^\intercal\Sigma_b^{-1}b_i\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[b_i^\intercal\left(\Sigma_b^{-1} + \frac{1}{\sigma^2}Z_i^\intercal Z_i\right)b_i - \frac{2}{\sigma^2}b_i^\intercal Z_i^\intercal r_i + \frac{1}{\sigma^2}r_i^\intercal r_i\right]\right\}
\end{aligned}
$$

Now we complete the square:

$x^\intercal Ax + x^\intercal b + c = (x-u)^\intercal A(x-u) + v$, where $u = -\frac{1}{2}A^{-1}b$ and $v = c - \frac{1}{4}b^\intercal A^{-1}b$.

We have

$$
\begin{aligned}
A &= \Sigma_b^{-1} + \frac{1}{\sigma^2}Z_i^\intercal Z_i \\
b &= -\frac{2}{\sigma^2}Z_i^\intercal r_i \\
\Rightarrow u &= -\frac{1}{2}\left(\Sigma_b^{-1} + \frac{1}{\sigma^2}Z_i^\intercal Z_i\right)^{-1}\left(-\frac{2}{\sigma^2}Z_i^\intercal r_i\right) \\
&= \frac{1}{\sigma^2}\left(\Sigma_b^{-1} + \frac{1}{\sigma^2}Z_i^\intercal Z_i\right)^{-1}Z_i^\intercal r_i
\end{aligned}
$$

We don't need to calculate v because it won't involve b_i.

Therefore,

$$\frac{f(y_i|b_i)g(b_i)}{h(y_i)} \propto \exp\left\{-\frac{1}{2}[b_i - u_i]^\intercal A_i [b_i - u_i]\right\}$$

which means that $b_i|y_i \sim \mathrm{N}(u_i, A_i^{-1})$ where u_i and A_i are defined above.

Now, to obtain the Q function from the log-likelihood, we'll use the following expectations:

$$\begin{aligned}
\underset{b_i|y_i}{\mathrm{E}} b_i^\intercal Z_i^\intercal r_i &= u_i^\intercal Z_i^\intercal r_i \\
\underset{b_i|y_i}{\mathrm{E}} b_i^\intercal Z_i^\intercal Z_i b_i &= \mathrm{tr}(Z_i^\intercal Z_i A_i^{-1}) + u_i^\intercal Z_i^\intercal Z_i u_i \\
&= \mathrm{tr}(Z_i A_i^{-1} Z_i^\intercal) + u_i^\intercal Z_i^\intercal Z_i u_i \\
\underset{b_i|y_i}{\mathrm{E}} b_i^\intercal \Sigma_b^{-1} b_i &= \mathrm{tr}(\Sigma_b^{-1} A_i^{-1}) + u_i^\intercal \Sigma_b^{-1} u_i
\end{aligned}$$

We have

$$\begin{aligned}
&Q(\beta, \Sigma_b, \sigma^2 | \beta^{(k)}, \Sigma_b^{(k)}, \sigma^{2(k)}) \\
&= \sum_{i=1}^n [-\frac{n_i}{2}\log(2\pi) - \frac{n_i}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} r_i^\intercal r_i + \frac{1}{\sigma^2} u_i^\intercal Z_i^\intercal r_i \\
&\quad - \frac{1}{2\sigma^2}\mathrm{tr}(Z_i A_i^{-1} Z_i^\intercal) - \frac{1}{2\sigma^2} u_i^\intercal Z_i^\intercal Z_i u_i \\
&\quad - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_b| \\
&\quad - \frac{1}{2}\mathrm{tr}(\Sigma_b^{-1} A_i^{-1}) - \frac{1}{2} u_i^\intercal \Sigma_b^{-1} u_i] \\
&= C - \frac{N}{2}\log(\sigma^2) - \frac{n}{2}\log|\Sigma_b| \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=1}^n r_i^\intercal r_i + \frac{1}{\sigma^2}\sum_{i=1}^n u_i^\intercal Z_i^\intercal r_i - \frac{1}{2\sigma^2}\sum_{i=1}^n u_i^\intercal Z_i^\intercal Z_i u_i \\
&\quad - \frac{1}{2\sigma^2}\sum_{i=1}^n \mathrm{tr}(Z_i A_i^{-1} Z_i^\intercal) \\
&\quad - \frac{1}{2}\sum_{i=1}^n \mathrm{tr}(\Sigma_b^{-1} A_i^{-1}) - \frac{1}{2}\sum_{i=1}^n u_i^\intercal \Sigma_b^{-1} u_i
\end{aligned}$$

where C is a constant, and $N = \sum_{i=1}^n n_i$.

M-Step Derivation

For the maximization step, we find $\beta, \Sigma_b, \sigma^2$ that maximize Q.

Setting the gradient of Q wrt β equal to 0 and solving, we have

$$\begin{aligned}
&\nabla_\beta Q = 0 \\
&\Rightarrow 0 = \frac{1}{\sigma^2}\sum_{i=1}^n X_i^\intercal (y_i - X_i\beta) - \frac{1}{\sigma^2}\sum_{i=1}^n X_i^\intercal Z_i u_i \\
&\Rightarrow 0 = \sum_{i=1}^n X_i^\intercal y_i - \sum_{i=1}^n (X_i^\intercal X_i)\beta - \sum_{i=1}^n X_i^\intercal Z_i u_i \\
&\Rightarrow \beta = \left(\sum_{i=1}^n X_i^\intercal X_i\right)^{-1} \sum_{i=1}^n X_i^\intercal (y_i - Z_i u_i) \\
&\qquad = \text{OLS estimate after subtracting predicted random effects from } y_i
\end{aligned}$$

Setting the gradient of Q wrt Σ_b equal to 0 and solving, we have

---

Reference for derivative-wrt-matrix identities:

https://www.ics.uci.edu/ welling/teaching/KernelsICS273B/MatrixCookBook.pdf

---

$$\begin{aligned}
&\nabla_{\Sigma_b} Q = 0 \\
&\Rightarrow 0 = -\frac{n}{2}\Sigma_b^{-1} + \frac{1}{2}\sum_{i=1}^n (\Sigma_b^{-1} A_i^{-1} \Sigma_b^{-1}) + \frac{1}{2}\sum_{i=1}^n (\Sigma_b^{-1} u_i u_i^\intercal \Sigma_b^{-1}) \\
&\Rightarrow n\Sigma_b \Sigma_b^{-1}\Sigma_b = \sum_{i=1}^n \Sigma_b \Sigma_b^{-1} u_i u_i^\intercal \Sigma_b^{-1}\Sigma_b + \sum_{i=1}^n \Sigma_b \Sigma_b^{-1} A_i^{-1} \Sigma_b^{-1} \Sigma_b \\
&\Rightarrow \Sigma_b = \frac{1}{n}\sum_{i=1}^n u_i u_i^\intercal + \frac{1}{n}\sum_{i=1}^n A_i^{-1} \\
&\qquad = \text{Covariance of conditional means of } b_i + \text{Mean of conditional covariances of } b_i
\end{aligned}$$

Finally, setting the gradient of Q wrt σ^2, we have

$$
\begin{aligned}
&\nabla_{\sigma^2} Q = 0 \\
&\quad \Rightarrow 0 = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n r_i^\intercal r_i - \frac{1}{2(\sigma^2)^2} 2\sum_{i=1}^n r_i^\intercal Z_i u_i + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n u_i^\intercal Z_i^\intercal Z_i u_i \\
&\qquad\qquad + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n \operatorname{tr}(Z_i A_i^{-1} Z_i^\intercal) \\
&\Rightarrow N\sigma^2 = \sum_{i=1}^n (r_i - Z_i u_i)^\intercal (r_i - Z_i u_i) + \sum_{i=1}^n \operatorname{tr}(Z_i A_i^{-1} Z_i^\intercal) \\
&\quad \Rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^n (y_i - X_i\beta - Z_i u_i)^\intercal (y_i - X_i\beta - Z_i u_i) + \frac{1}{N}\sum_{i=1}^n \operatorname{tr}(Z_i A_i^{-1} Z_i^\intercal) \\
&\qquad = \text{Variance from deviations of } y_i \text{ from conditional mean} \\
&\qquad\qquad + \text{Variance from random effects}
\end{aligned}
$$

where we substitute β with the estimate derived above, $\left(\sum_{i=1}^n X_i^\intercal X_i\right)^{-1} \sum_{i=1}^n X_i^\intercal (y_i - Z_i u_i)$.

Algorithm

Putting this all together, an EM algorithm for estimating β, Σ_b, and σ^2 is

1. Set $\beta^{(0)}$ (e.g. with OLS estimates), $\Sigma_b^{(0)}$ (e.g. with I_q), and $\sigma^{2(0)}$ (e.g. with OLS estimate).
2. At iteration k (E step), given current estimates $\beta^{(k)}$, $\Sigma_b^{(k)}$, and $\sigma^{2(k)}$,
   - Calculate $A_i^{-1} = \left(\Sigma_b^{-1(k)} + \frac{1}{\sigma^{2(k)}} Z_i^\intercal Z_i\right)^{-1}$, $i = 1, \ldots, n$.
   - Calculate $u_i = \frac{1}{\sigma^{2(k)}} A_i^{-1} Z_i^\intercal (y_i - X_i \beta^{(k)})$, $i = 1, \ldots, n$.
3. (Iteration k, M step) Calculate new estimates using $u_1, \ldots, u_n$ and $A_1^{-1}, \ldots, A_n^{-1}$:

$$
\begin{aligned}
\beta^{(k+1)} &= \left(\sum_{i=1}^n X_i^\intercal X_i\right)^{-1} \sum_{i=1}^n X_i^\intercal (y_i - Z_i u_i) \\
\Sigma_b^{(k+1)} &= \frac{1}{n}\sum_{i=1}^n u_i u_i^\intercal + \frac{1}{n}\sum_{i=1}^n A_i^{-1} \\
\sigma^{2(k)} &= \frac{1}{N}\sum_{i=1}^n (y_i - X_i\beta^{(k+1)} - Z_i u_i)^\intercal (y_i - X_i\beta^{(k+1)} - Z_i u_i) + \frac{1}{N}\sum_{i=1}^n \operatorname{tr}(Z_i A_i^{-1} Z_i^\intercal)
\end{aligned}
$$

4. Continue with iterations of steps 2–3 until convergence.