Mixed Model Prediction: Part III

Jiming Jiang and Thuan Nguyen

University of California, Davis, USA

and Oregon Health & Science University, USA

LMM prediction

- Mixed model prediction (MMP) has a fairly long history starting with Henderson's early work in animal breeding (Henderson 1948).
  The field has since flourished, thanks to its broad applications in various fields. Examples: animal and plant genetics, personalized medicine, business and economics, education, surveys, ...
  Example 1 (IQ test). The following (hypothetical) example was given by Mood *et al.* (1974, p. 370). Suppose that it is known that the IQ of students in a particular age group are normally distributed with mean 100 and s.d. 15.
  It is also known that, for a given student, the test scores are normally distributed with mean equal to the student's IQ and s.d. equal to 5.

- Suppose that a student just took an IQ test and scored 130.

  What is the best estimate (or prediction) of the student's IQ?

  The answer is not 130.

  If I had not told you all of the stories about the IQ, the best prediction would be 130.

  Before we go any further, can you speculate if the best prediction is greater than 130, or less than 130?

- Best prediction (BP): In the sense of mean squared prediction error (MSPE), the best predictor of η, an unobserved random variable, based on y, an observed random variable, is $\tilde{\eta} = \mathrm{E}(\eta|y)$.

  A formula for conditional expectation: If

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right],$$

  then

$$\mathrm{E}(\xi_1|\xi_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\xi_2 - \mu_2).$$

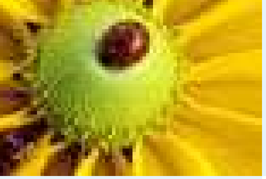
- Back to the IQ problem, let $\xi_1 = \mathrm{IQ}$, $\xi_2 =$score. We know $\mathrm{IQ} \sim N(100, 15^2)$, and score$|\mathrm{IQ} \sim N(\mathrm{IQ}, 5^2)$.

  Thus, it is easy to derive: $\mu_1 = \mu_2 = 100$, $\Sigma_{11} = \Sigma_{12} = \Sigma_{21} = 15^2$, and $\Sigma_{22} = 15^2 + 5^2$.

  Thus, the BP of the person's IQ is

$$\tilde{\mathrm{IQ}} = 100 + \frac{15^2}{15^2 + 5^2}(130 - 100) = 127.$$

  Cool?



- In general, under a Gaussian mixed model, we have

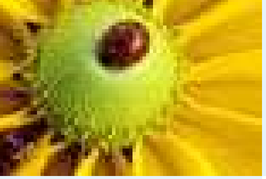
$$\begin{pmatrix} \alpha \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} G & GZ' \\ ZG & V \end{pmatrix}\right].$$

Thus, according to the normal prediction theory, we have

(1) $$\mathrm{E}(\alpha|y) = GZ'V^{-1}(y - X\beta).$$

Now suppose that we are interested in a mixed effect, $\xi = b'\beta + a'\alpha$, where a, b are constant (known) vectors. From (3), we get

$$\begin{aligned} \mathrm{E}(\xi|y) &= b'\beta + a'\mathrm{E}(\alpha|y) \\ &= b'\beta + a'GZ'V^{-1}(y - X\beta). \end{aligned} \quad (2)$$



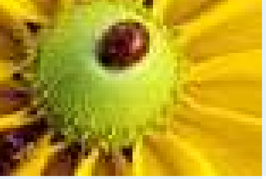
- The expressions (3) or (4) are the (theoretical) BPs of α or ξ, respectively, but they are usually not computable, because β and V are unknown.

  To go one step further, let us assume that V is known, for now. Then, it is natural to replace the unknown β by its MLE,

  $$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{3}$$

  (5) is also known as the best linear unbiased estimator (BLUE) of β, and the definition means what it means even without normality.

  Once the β in the BPs, (3) or (4), is replaced by its BLUE, the result is called best linear unbiased predictor, or BLUP.



- Alternative derivations of BLUP:

  Henderson (1950) gave the first derivation of BLUP using the idea of what he called “maximum likelihood estimates” of fixed and random effects.

  Harville (1990) showed that BLUP is actually, well, BLUP, that is, it minimizes the MSPE among all linear and unbiased predictors of ξ.

  Jiang (1997b) derived BLUP as the BP based on error contrasts. For example, the BLUP of α is equal to $\mathrm{E}(\alpha|A'y)$ for any matrix A satisfying the REML condition (2).

- Empirical BLUP (EBLUP):

  The BLUP typically still involves unknown parameters, namely, θ.

  Thus, as the final step, we replace θ by $\hat{\theta}$, a consistent estimator. The result is called empirical BLUP, or EBLUP.

  In conclusion, we have

$$
\begin{aligned}
&(4) \qquad & \text{BLUP} &= \text{BP with } \beta \text{ replaced by } (5); \\
&(5) \qquad & \text{EBLUP} &= \text{BLUP with } \theta \text{ replaced by } \hat{\theta}.
\end{aligned}
$$

  For example, $\hat{\theta}$ may be the REML or ML estimators of θ.

- Example 2: Robinson (1991) used the following hypothetical example to illustrate the BLUP.

  Consider the following data from lactation yields of dairy cows: Row 1: Herd; Row 2: Sire; Row 3: Yield.

1	1	2	2	2	3	3	3	3
A	D	B	D	D	C	C	D	D
110	100	110	100	100	110	110	100	100

  Standard LMM assumptions $y = X\beta + Z\alpha + \epsilon$, ...

  $R = I, G = 0.1 * I$.

- The BLUE is computed as

$$\tilde{\beta} = (105.64, 104.28, 105.46)'.$$

The BLUP of α, the sire effects, is computed as

$$\tilde{\alpha} = (0.40, 0.52, 0.76, -1.67)'.$$

Interpretation?

Data again (overall mean $= 104.44$):

1	1	2	2	2	3	3	3	3
A	D	B	D	D	C	C	D	D
110	100	110	100	100	110	110	100	100

GLMM prediction

- 1. Joint inference about fixed and random effects

  Henderson initially derived the BLUP as joint estimation of fixed and random effects

  The idea can be extended as that of *maximum a posterior*

  As noted, the BLUP may be derived as a way of joint estimation of fixed and random effects.

  Let y be a vector of observations, γ a vector of unobserved “random variables”, and ψ a vector of parameters.

- Here, γ may be a vector of random effects, or a vector of fixed and random effects; correspondingly, ψ may be a vector of fixed effects and variance components, or a vector of variance components only.

  Let $p(y, \gamma|\psi)$ denote the joint pdf of y and γ, with respect to a σ-finite measure, ν, given that ψ is the true parameter vector.

  Note that, in case that γ involves the fixed effects, one may need to define what is meant by the distribution of γ, perhaps, under a Bayesian framework, but this is not important, at least at this point.

- What is important is the following relationship:

$$p(y, \gamma|\psi) = p(y|\psi)p(\gamma|y, \psi), \quad (6)$$

where $p(y|\psi)$ and $p(\gamma|y, \psi)$ denote the marginal pdf of y and conditional pdf of γ given y, respectively, given that ψ is the true parameter vector.
Using a Bayesian term, $p(\gamma|y, \psi)$ is called *a posterior*.
Henderson's original idea (Henderson 1950) was to find $\hat{\gamma} = \hat{\gamma}(y, \psi)$ that maximizes the left side of (1). From the right side of the same equation, this is equivalent to finding $\hat{\gamma}$ that maximizes $p(\gamma|y, \psi)$, the posterior.
In other words, the BLUP may be regarded as a maximum a posterior (MPE) estimator of γ, and this concept is not restricted to linear models.

- To be specific, let $\gamma = (\beta', \alpha')'$, where β and α are vectors of fixed and random effects, respectively.

  The MPE of β and α are typically obtained by solving a system of equations that equal the derivatives of $l(y, \gamma|\psi) = \log\{p(y, \gamma|\psi)\}$ to zero, that is,

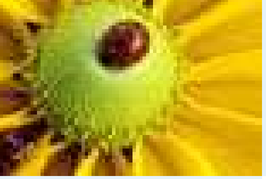
$$\begin{aligned} \frac{\partial l}{\partial \beta} &= 0, \\ \frac{\partial l}{\partial \alpha} &= 0, \end{aligned}$$

  In practice, there are often a large number of random effects involved in a GLMM.

- For example, the number of female and male random effects involved in the salamander problem is 80 (or 120 under the pooled data model); the number of HSA-specific random effects involved in the BRFSS problem is 118.

  Standard methods of solving nonlinear systems, such as Newton–Raphson (N-R), may be inefficient and extremely slow when the dimension of the solution is high.

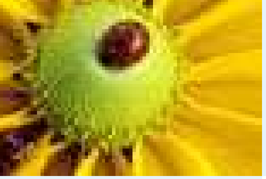
  Jiang (2000) proposed a nonlinear Gauss–Seidel algorithm (NLGSA) for computing the MPEs, and proved global convergence of the algorithm.



- Example 3. Consider a simplified version of the salamander problem. Namely, given the random effects $u_1, \dots, u_m$ and $v_1, \dots, v_n$, binary responses y_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n$ are conditionally independent such that, with $p_{ij} = \mathrm{P}(y_{ij} = 1 | u, v)$, one has $\mathrm{logit}(p_{ij}) = \mu + u_i + v_j$, where where μ is an unknown parameter.

  Assume that the random effects are independent with $u_i \sim N(0, \sigma^2)$ and $v_j \sim N(0, \tau^2)$.

  To illustrate the NLGSA, assume, for simplicity, that μ, σ^2, and τ^2 are known.



- It can be shown that maximum a posterior, which is equivalent to maximizing the joint pdf $f(y,u,v)$ with respect to u, v, leads to the following system of nonlinear equations given μ:

$$
\begin{aligned}
(7) \quad \frac{u_i}{\sigma^2} + \sum_{j=1}^{n} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} &= y_{i\cdot}, \quad 1 \le i \le m, \\
(8) \quad \frac{v_j}{\tau^2} + \sum_{i=1}^{m} \frac{\exp(\mu + u_i + v_j)}{1 + \exp(\mu + u_i + v_j)} &= y_{\cdot j}, \quad 1 \le j \le n,
\end{aligned}
$$

where $y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$ and $y_{\cdot j} = \sum_{i=1}^{m} y_{ij}$.

- Note that, given the v_js, each equation (2) univariate, which can be easily solved. A similar observation is made regarding (3). This motivates the following algorithm:

  Starting with initial values $v_j^{(0)}$, $1 \leq j \leq n$, solve (2) with $v_j^{(0)}$ in place of v_j, $1 \leq j \leq n$ to get $u_i^{(1)}$, $1 \leq i \leq m$;

  then, solve (3) with $u_i^{(1)}$ in place of u_i, $1 \leq i \leq m$ to get $v_j^{(1)}$, $1 \leq j \leq n$;

  and so on ...

- Example 4. Recall the BRFSS problem of estimating the proportion of women having had mammography.

  The mixed logistic model can be expressed as

$$\begin{aligned} \mathrm{logit}(p) &= \beta_0 + \beta_1 * \mathrm{age} + \beta_2 * \mathrm{age}^2 + \beta_3 * \mathrm{Race} \\ &\quad + \beta_4 * \mathrm{Edu} + \mathrm{HSA\ effect}. \end{aligned}$$

  The MPE of the fixed effects are $\hat{\beta}_0 = -0.421$, $\hat{\beta}_1 = 0.390$, $\hat{\beta}_2 = -0.047$, $\hat{\beta}_3 = -0.175$, and $\hat{\beta}_4 = 2.155$.

- Furthermore, the variance of the HSA effects can be estimated by

$$\hat{\sigma}^2 = \frac{1}{118}\sum_{i=1}^{118}\hat{v}_i^2,$$

where $\hat{v}_i$ is the MPE of the ith HSA effect. This gives $\hat{\sigma} = 0.042$.

Note that the variance components, θ, are involved in the MP equations. So, technically, to obtain the MPE one needs to know θ.

However, Jiang *et al.* (2001) showed that the consistency property of MPE is not affected by at which θ the MPEs are evaluated. In other words, the MP equations can be solved with whatever (reasonable guess) of θ, and the resulting MPE still have good behavior asymptotically, under suitable conditions.

- 2. Empirical best prediction (EBP)

Consider a special form of GLMM. Suppose that, conditional on a vector of random effects, $\alpha_i = (\alpha_{ij})_{1 \le j \le r}$, responses $y_{ij}, 1 \le j \le n_i$ are independent with density

$$f(y_{ij}|\alpha_i) \quad = \quad \exp\left[\left(\frac{a_{ij}}{\phi}\right)\{y_{ij}\xi_{ij} - b(\xi_{ij})\} + c\left(y_{ij}, \frac{\phi}{a_{ij}}\right)\right],$$

where $b(\cdot)$ and $c(\cdot, \cdot)$ are functions associated with the exponential family;

ϕ is a dispersion parameter, and a_{ij} is a weight such that $a_{ij} = 1$ for ungrouped data; $a_{ij} = l_{ij}$ for grouped data when the average is considered as response and l_{ij} is the group size; and $a_{ij} = l_{ij}^{-1}$ when the sum of individual responses is considered.

prediction

prediction

- Furthermore, ξ_{ij} is associated with a linear function

$$\eta_{ij} = x_{ij}'\beta + z_{ij}'\alpha_i$$

through a link function $g(\cdot)$; that is, $g(\xi_{ij}) = \eta_{ij}$.

In the case of a canonical link, we have $\xi_{ij} = \eta_{ij}$.

Finally, suppose that $v_1, \dots, v_m$ are independent with density $f_\theta(\cdot)$, where θ is a vector of variance components.

Consider the problem of predicting a mixed effect of the following form,

$$\zeta = \zeta(\beta, \alpha_S),$$

where S is a subset of $\{1, \dots, m\}$, and $\alpha_S = (\alpha_i)_{i \in S}$.

- Let $y_S = (y_i)_{i \in S}$, where $y_i = (y_{ij})_{1 \le j \le n_i}$ and $y_{S-} = (y_i)_{i \notin S}$.

  Then, the BP of ζ is given by

$$
\begin{aligned}
\tilde{\zeta} &= \mathrm{E}(\zeta|y) \\
&= \mathrm{E}(\zeta(\beta, \alpha_S)|y_S) \\
&= \frac{\int \zeta(\beta, \alpha_S) f(y_S|\alpha_S) f_\theta(\alpha_S) d\alpha_S}{\int f(y_S|\alpha_S) f_\theta(\alpha_S) d\alpha_S} \\
&= \frac{\int \zeta(\beta, \alpha_S) \exp\{\phi^{-1} \sum_{i \in S} s_i(\beta, \alpha_i)\} \prod_{i \in S} f_\theta(\alpha_i) \prod_{i \in S} d\alpha_i}{\prod_{i \in S} \int \exp\{\phi^{-1} s_i(\beta, v)\} f_\theta(v) dv},
\end{aligned}
$$

  where $s_i(\beta, v) = \sum_{j=1}^{n_i} a_{ij}[y_{ij} h(x'_{ij}\beta + z'_{ij}v) - b\{h(x_{ij}\beta + z_{ij}v)\}]$.

  The EBP is the BP with the unknown parameters replaced by their (consistent) estimators.