Mixed-Models

version 30 October 2011

Mixed models

- Mixed models estimate a vector β of fixed effects and one (or more) vectors $\mathbf{u}$ of random effects
  - Both fixed and random effects models always include a vector $\mathbf{e}$ of residuals. These are random effects as well
- We speak about estimating fixed effects (BLUE)
- And **predicting** random effects
  - BLUP = best linear unbiased predictor

Random effects

- Why use random effects?
- Biological motivation: We consider our sample of values as being drawn from some much larger universe of interest
- Statistical motivation: treating effects as random generally uses much fewer degrees of freedom!
- Nuisance parameters often treated as random

Example

- Suppose you are measuring individuals of interest from different locations, and/or different years
- Year, location, and year x location effects may be important
- However, treating all these as fixed uses up degrees of freedom
- Conversely, simply ignoring these gives a more complicated residual error structure

Example (cont)

- Suppose you have measured 3 individuals from location 1, 2 from location 2, and 4 from location 3.
- Let u_1, u_2, and u_3 denote the location values for this particular sample.
- Ignoring these effects creates a covariance between residuals from the same location (as they have u_i in common)
  - A common reason to include random effects is to reduce the residual error structure back to the OLS form, σ_e^2 I.

Focusing just on the random effects in this model, we have **y = Zu + e**

$$\mathbf{y}_{9\times 1} = \begin{pmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \\ y_{3,4} \end{pmatrix}, \quad \mathbf{u}_{3\times 1} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \mathbf{Z}_{9\times 3} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Z is called the **incident matrix** for the random effect

While **Z** is critical (the analogue of the design matrix **X** associated With fixed effects), it is NOT sufficient to characterize the random effects. Most importantly, we need their **covariance structure**: (the variance-covariance matrix)

Example (cont)

Suppose we assume that the location effects are drawn from the same distribution (and hence have the same variance σ_u^2) and are uncorrelated.

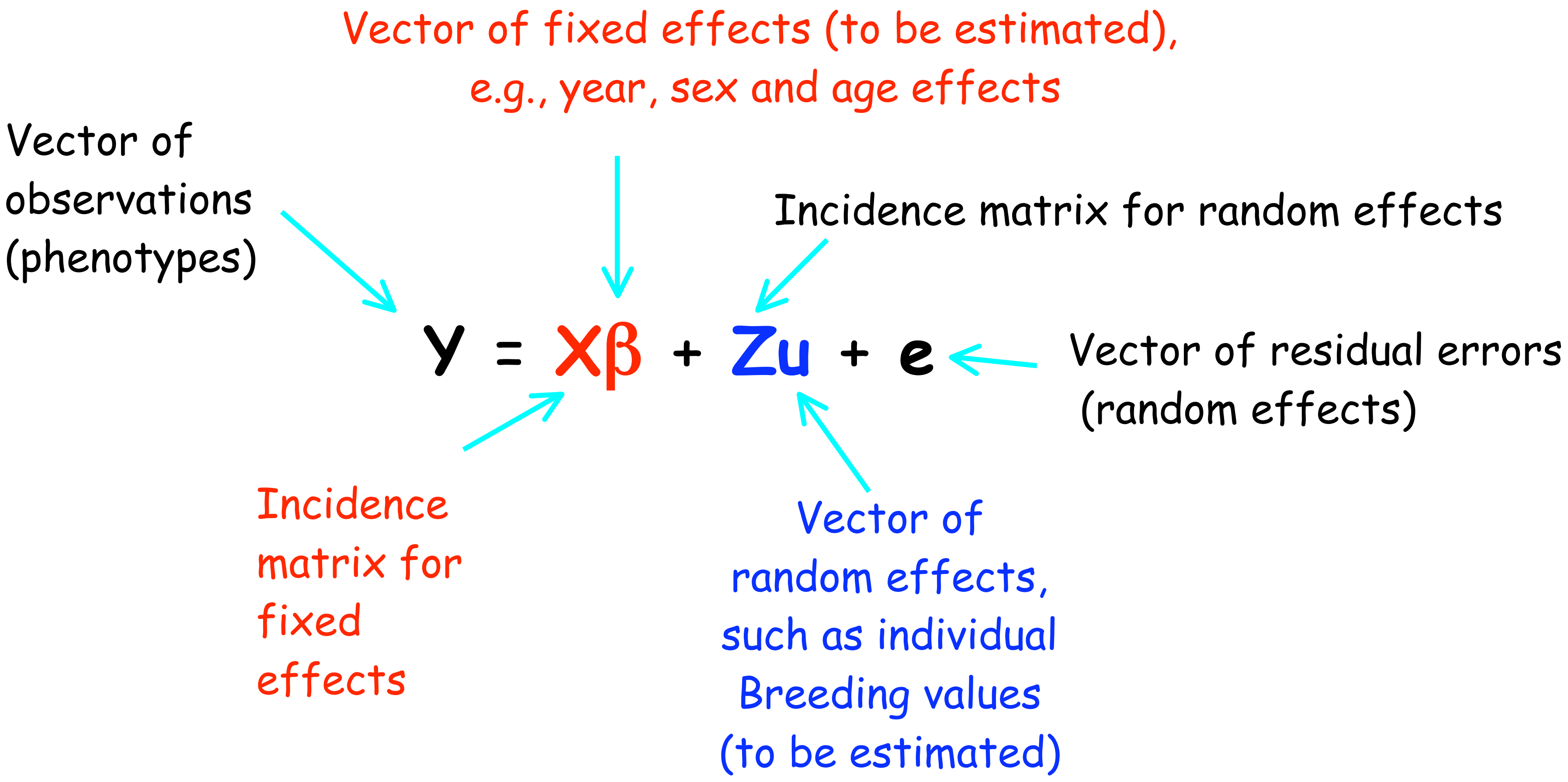
Then $\mathbf{R} = \text{cov}(\mathbf{u}) = \sigma_u^2\, \mathbf{I}$

The resulting covariance matrix for the y = Zu + e (making The OLS assumption for e) becomes

$$\text{Var}(\mathbf{y}) = \mathbf{Z}^T\mathbf{R}\mathbf{Z} + \sigma_s^2\, \mathbf{I} \;\; = \mathbf{Z}^T(\sigma_u^2\, \mathbf{I})\mathbf{Z} + \sigma_s^2\, \mathbf{I} = \sigma_u^2\, \mathbf{Z}^T\, \mathbf{Z} + \sigma_s^2\, \mathbf{I}$$

The covariance among observations for the same location is fully accounted for by the resulting covariance matrix

The general mixed model



Vector of fixed effects (to be estimated),
e.g., year, sex and age effects

Vector of observations (phenotypes)

Incidence matrix for random effects

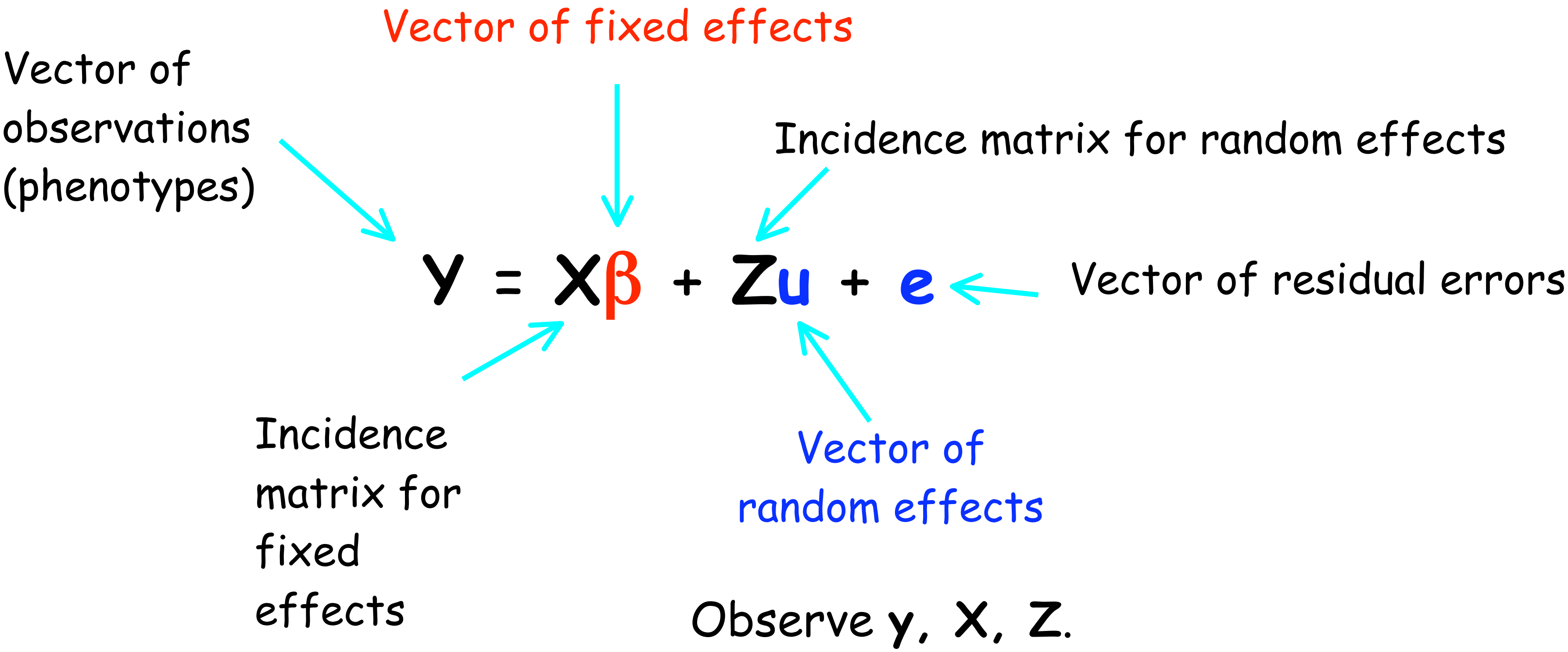
$$y = X\beta + Zu + e$$

Vector of residual errors (random effects)

Incidence matrix for fixed effects

Vector of random effects, such as individual Breeding values (to be estimated)

The general mixed model



Vector of fixed effects

Vector of observations (phenotypes)

Incidence matrix for random effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Vector of residual errors

Incidence matrix for fixed effects

Vector of random effects

Observe $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z}$.

Estimate fixed effects β

Estimate random effects $\mathbf{u}$, $\mathbf{e}$

Means & Variances for $y = X\beta + Zu + e$

Means: $E(u) = E(e) = 0$, $E(y) = X\beta$

Variances:

Let R be the covariance matrix for the residuals. We typically assume $R = \sigma^2_e * I$

Let G be the covariance matrix for the breeding values (the vector **u**)

The covariance matrix for y becomes

$$V = ZGZ^T + R$$

Effects of model misspecification

Suppose we simply used a General Linear model (only fixed effects) for this example?

Here $y = X\beta + e^*$, where $e^* \sim MVN(0,V)$, implying $Y \sim MVN(X\beta,V)$

The effect of using a mixed model is that it partitions the residual e^* as $e^* = Zu + e$

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe **y**, **X**, and **Z**

β, u, R, and G are generally unknown

Two complementary estimation issues

(i) Estimation of β and **u**

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $V = ZGZ^T + R$

Henderson's Mixed Model Equations

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$, $\mathbf{u} \sim (0, G)$, $\mathbf{e} \sim (0, \mathbf{R})$, $\mathrm{cov}(\mathbf{u},\mathbf{e}) = 0$,

If X is n x p and Z is n x q

$$\begin{array}{cc} p \times p & p \times q \\ \multicolumn{2}{c}{\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}} \\ & q \times q \end{array} \begin{pmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

The whole matrix is (p+q) x (p+q)

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$$

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)$$

Inversion of an n x n matrix

Standard Errors

A relatively straightforward extension of Henderson's mixed-model equations provides estimates of the standard errors of the fixed and random effects. Let the inverse of the leftmost matrix in Equation 26.5 be

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{26.6}$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $p \times p$, $p \times q$, and $q \times q$ submatrices. Using this notation, Henderson (1975) showed that the sampling covariance matrix for the BLUE of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}\,) = \mathbf{C}_{11} \tag{26.7a}$$

that the sampling covariance matrix of the prediction errors $(\widehat{\mathbf{u}} - \mathbf{u})$ is given by

$$\boldsymbol{\sigma}(\,\widehat{\mathbf{u}} - \mathbf{u}\,) = \mathbf{C}_{22} \tag{26.7b}$$

and that the sampling covariance of estimated effects and prediction errors is given by

$$\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}} - \mathbf{u}\,) = \mathbf{C}_{12} \tag{26.7c}$$

(We consider $\widehat{\mathbf{u}} - \mathbf{u}$ rather than $\widehat{\mathbf{u}}$ as the latter includes variance from both the prediction error and the random effects $\mathbf{u}$ themselves.)

The Animal Model, $y_i = \mu + a_i + e_i$

Here, the individual is the unit of analysis, with y_i the phenotypic value of the individual and a_i its BV

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \mu, \qquad \mathbf{u} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} \qquad \mathbf{G} = \sigma^2_A\,\mathbf{A},$$

Where the additive genetic relationship matrix A is given by

$A_{ij} = 2\theta_{ij}$ namely twice the coefficient of coancestry

Assume $R = \sigma^2_e * I$, so that $R^{-1} = 1/(\sigma^2_e) * I$.

Likewise, $G = \sigma^2_A * A$, so that $G^{-1} = 1/(\sigma^2_A) * A^{-1}$.

Henderson's mixed model equation here becomes

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

here $\lambda = \sigma^2_e \,/\, \sigma^2_A = (1-h^2)/h^2$

This reduces to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \sum^n y_i \\ \mathbf{y} \end{pmatrix}$$

Estimation of R and G

The second estimation issue the covariance matrix for residuals R and for breeding values G

As we have seen, both matrices have the form $\sigma^2 * B$, where the variance σ^2 is unknown, but B is known

For example, for residuals, $R = \sigma^2_e * I$

For breeding values, $G = \sigma^2_A * A$, where A is given from the pedigree

REML Variance Component Estimation

REML = Restricted Maximum Likelihood.

Standard ML variance estimation assumes fixed factors are known without error. Results in downward bias in variance estimates

REML maximizes that portion of the likelihood that does not depend on fixed effects

Basic idea: Use a transformation to remove fixed effect, then perform ML on this transformed vector

Simple variance estimate under ML vs. REML

$$\text{ML} = \frac{1}{n}\sum_{i+1}^{n}(x-\overline{x})^2, \quad \text{REML} = \frac{1}{n-1}\sum_{i+1}^{n}(x-\overline{x})^2$$

REML adjusts for the
estimated fixed
effect,
in this case, the mean

With balanced design, ANOVA variance estimates are equivalent to REML variance estimates