

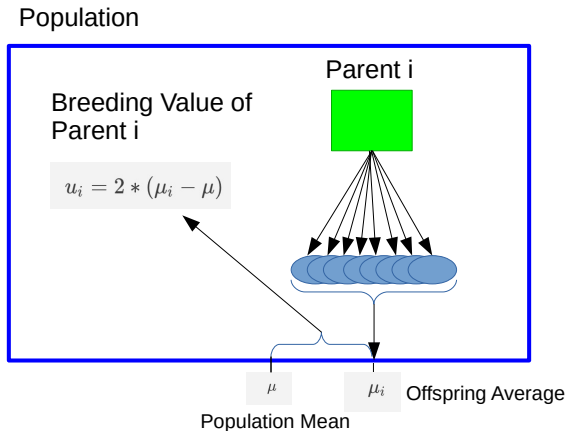
Prediction of Breeding Values

Peter von Rohr

2022-05-11

What are breeding values

Definition: two times difference between offspring of a given parent from population mean



Practical Considerations

- ▶ Definition of breeding value is based on biological fact that parent passes half of its alleles to offspring
- ▶ In practice, definition cannot be used
 - ▶ most parents do not have enough offspring
 - ▶ breeding values are needed before animals have offspring
 - ▶ different environmental factors not considered

Solution

- ▶ Use genetic model to predict breeding values based on phenotypic observations
- ▶ Genetic model decomposes phenotypic observation (y_i) in different components

$$y_i = \mu + u_i + d_i + i_i + e_i$$

where μ is the general mean, u_i the breeding value, d_i the dominance deviation, i_i the epistasis effect and e_i the random error term.

Solution II

- ▶ For predicting breeding values d_i and i_i are often ignored, leading to a simplified version of the genetic model

$$y_i = \mu + u_i + e_i$$

- ▶ Expected values and variance-covariance matrix

$$E \begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \\ 0 \end{bmatrix}$$
$$\text{var} \begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} = \begin{bmatrix} \sigma_y^2 & \sigma_u^2 & \sigma_e^2 \\ \sigma_u^2 & \sigma_u^2 & 0 \\ \sigma_e^2 & 0 & \sigma_e^2 \end{bmatrix}$$

How to Predict Breeding Values

- ▶ Predicted breeding values (\hat{u}) are a function of the observed phenotypic data (y)

$$\rightarrow \hat{u} = f(y)$$

- ▶ What should $f()$ look like?
- ▶ Goal: Maximize improvement of offspring generation over parents

$\rightarrow \hat{u}$ should be conditional expected value of true breeding value u given y :

$$\hat{u} = E(u|y)$$

Derivation

- ▶ Assume: multivariate normality of u and y and $E(u) = 0$, then

$$\begin{aligned}\hat{u} &= E(u|y) = E(u) + \text{cov}(u, y^T) * \text{var}(y)^{-1} * (y - E(y)) \\ &= E(u|y) = \text{cov}(u, y^T) * \text{var}(y)^{-1} * (y - E(y))\end{aligned}$$

- ▶ \hat{u} consists of two parts
 1. $(y - E(y))$: phenotypic observations corrected for environmental effects
 2. $\text{cov}(u, y^T) * \text{var}(y)^{-1}$: weighting factor of corrected observation

Unbiasedness

- ▶ Expected value ($E(\hat{u})$)

$$\begin{aligned} E(\hat{u}) &= E(\text{cov}(u, y^T) * \text{var}(y)^{-1} * (y - E(y))) \\ &= \text{cov}(u, y^T) * \text{var}(y)^{-1} * E(y - E(y)) \\ &= \text{cov}(u, y^T) * \text{var}(y)^{-1} * (E(y) - E(y)) = 0 \end{aligned}$$

- ▶ With $E(u) = 0$, it follows $E(\hat{u}) = E(u) = 0$

Variance

- ▶ $\text{var}(\hat{u})$ and $\text{cov}(u, \hat{u})$ important for quality of prediction

$$\begin{aligned}\text{var}(\hat{u}) &= \text{var}(\text{cov}(u, y^T) * \text{var}(y)^{-1} * (y - E(y))) \\ &= \text{cov}(u, y^T) * \text{var}(y)^{-1} * \text{var}(y - E(y)) \\ &\quad * \text{var}(y)^{-1} * \text{cov}(y, u^T) \\ &= \text{cov}(u, y^T) * \text{var}(y)^{-1} * \text{cov}(y, u^T)\end{aligned}$$

$$\begin{aligned}\text{cov}(u, \hat{u}) &= \text{cov}(u, (\text{cov}(u, y^T) * \text{var}(y)^{-1} * (y - E(y))))^T) \\ &= \text{cov}(u, (y - E(y))^T) * \text{var}(y)^{-1} * \text{cov}(y, u^T) \\ &= \text{cov}(u, y^T) * \text{var}(y)^{-1} * \text{cov}(y, u^T) = \text{var}(\hat{u})\end{aligned}$$

Accuracy

- ▶ Measured by $r_{u,\hat{u}}$
- ▶ Recall $\text{cov}(u, \hat{u}) = \text{var}(\hat{u})$

$$\begin{aligned}r_{u,\hat{u}} &= \frac{\text{cov}(u, \hat{u})}{\sqrt{\text{var}(u) * \text{var}(\hat{u})}} \\ &= \sqrt{\frac{\text{var}(\hat{u})}{\text{var}(u)}}\end{aligned}$$

- ▶ Reliability (“Bestimmtheitsmass”): $B = r_{u,\hat{u}}^2$

Prediction Error Variance (PEV)

- ▶ Variability of prediction error: $u - \hat{u}$

$$\begin{aligned} \text{var}(u - \hat{u}) &= \text{var}(u) - 2\text{cov}(u, \hat{u}) + \text{var}(\hat{u}) = \text{var}(u) - \text{var}(\hat{u}) \\ &= \text{var}(u) * \left[1 - \frac{\text{var}(\hat{u})}{\text{var}(u)} \right] \\ &= \text{var}(u) * \left[1 - r_{u, \hat{u}}^2 \right] \end{aligned}$$

- ▶ Obtained from coefficient matrix of mixed model equations
- ▶ Used to compute reliability

Conditional Density

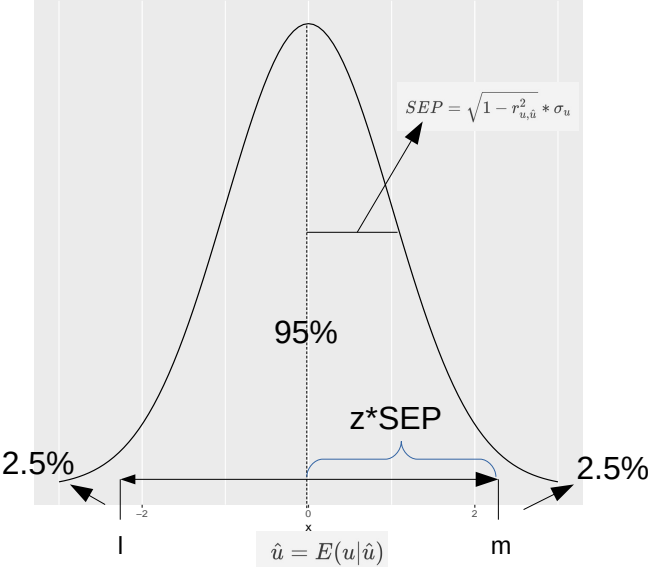
- ▶ Assessment of risk when using animals with predicted breeding values with different reliabilities quantified by $f(u|\hat{u})$
- ▶ Multivariate normal density with mean $E(u|\hat{u})$ and variance $var(u|\hat{u})$

$$\begin{aligned}E(u|\hat{u}) &= E(u) + cov(u, \hat{u}^T) * var(\hat{u})^{-1} * (\hat{u} - E(\hat{u})) = \hat{u} \\var(u|\hat{u}) &= var(u) - cov(u, \hat{u}^T) * var(\hat{u})^{-1} * cov(\hat{u}, u^T) \\&= var(u) * \left[1 - \frac{cov(u, \hat{u}^T)^2}{var(u) * var(\hat{u})} \right] \\&= var(u) * \left[1 - r_{u, \hat{u}}^2 \right]\end{aligned}$$

Confidence Intervals (CI)

- ▶ Assume an error level α , this results in $100 * (1 - \alpha)\%$ -CI
- ▶ Typical values of α 0.05 or 0.01
- ▶ With $\alpha = 0.05$, the 95%-CI gives interval around mean which covers a surface of 0.95

CI-Plot



CI Limits

- ▶ lower limit l and upper limit m are given by

$$\begin{aligned}l &= \hat{u} - z * SEP \\m &= \hat{u} + z * SEP\end{aligned}\tag{1}$$

- ▶ z corresponds to quantile value to cover a surface of $(1 - \alpha)$
- ▶ Use R-function `qnorm()` to get value of z

Linear Mixed Effects Model

- ▶ Use more realistic model for prediction of breeding values

$$y = Xb + Zu + e$$

where

- y vector of length n with observations
- b vector of length p with fixed effects
- u vector of length q with random breeding values
- e vector of length n with random error terms
- X $n \times p$ incidence matrix
- Z $n \times q$ incidence matrix

Expected Values and Variances

$$E \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \end{bmatrix}$$

$$\text{var} \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} ZGZ^T + R & ZG & 0 \\ & GZ^T & G & 0 \\ & & 0 & 0 & R \end{bmatrix}$$

Solutions

- ▶ Same as for simple model

$$\hat{u} = E(u|y) = GZ^T V^{-1}(y - X\hat{b})$$

with

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

corresponding to the general least squares solution of b

Problem

- ▶ Solution for \hat{u} contains V^{-1} which is large and difficult to compute
- ▶ Use mixed model equations

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

Sire Model

$$y = Xb + Zs + e$$

where s is a vector of length q_s with all sire effects.

$$\text{var}(s) = A_s * \sigma_s^2$$

where A_s : numerator relationship considering only sires

Animal Model

$$y = Xb + Za + e$$

where a is a vector of length q_a containing the breeding values

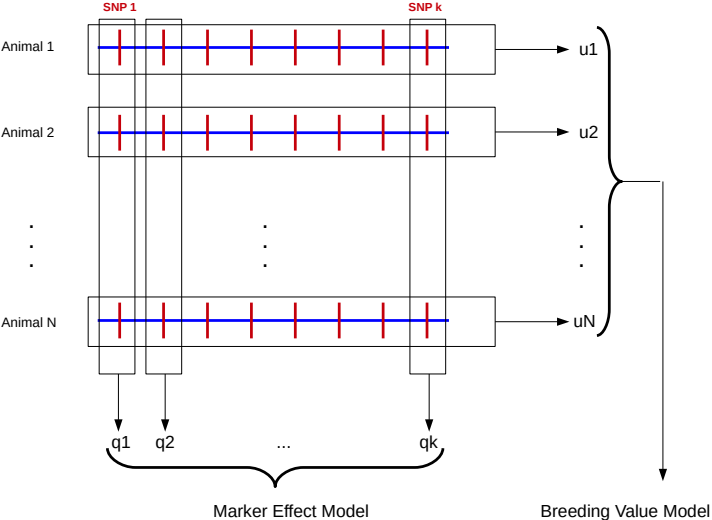
$$\text{var}(a) = A\sigma_a^2$$

where A is the numerator relationship matrix

Genomic BLUP (GBLUP)

1. marker-effect models: SNP-loci as random effects (MEM)
2. breeding value based models: genomic breeding values as random effects (BVM)

MEM and BVM



Marker Effect Model (MEM)

- ▶ Marker effects (a) as random in a linear mixed effects model

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

- ▶ Solution of marker effects via mixed model equations

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} M \\ M^T R^{-1} X & M^T R^{-1} M + I * \sigma_a^{-2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ M^T R^{-1} y \end{bmatrix}$$

where σ_a^2 is the SNP-effect variance component.

- ▶ Genomic breeding value for animal i is computed as sum over appropriate values of \hat{a} given by genotype of animal i

Breeding Value Based Model

- ▶ Genomic breeding values (u) as random effects in linear mixed effects model

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

- ▶ Solutions

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} W \\ W^T R^{-1} X & W^T R^{-1} W + G^{-1} * \sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ W^T R^{-1} y \end{bmatrix}$$

- ▶ Genomic breeding values correspond to solutions for \hat{u}

How Does GBLUP Work

$$\begin{bmatrix} X^T X & X^T Z & 0 \\ Z^T X & Z^T Z + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \\ 0 \end{bmatrix}$$

- ▶ $G^{(11)}$: animals with phenotypic observations
- ▶ $G^{(22)}$: animals without genotypic observations

$$\hat{g}_2 = - \left(G^{(22)} \right)^{-1} G^{21} \hat{g}_1$$