Prediction of Breeding Values

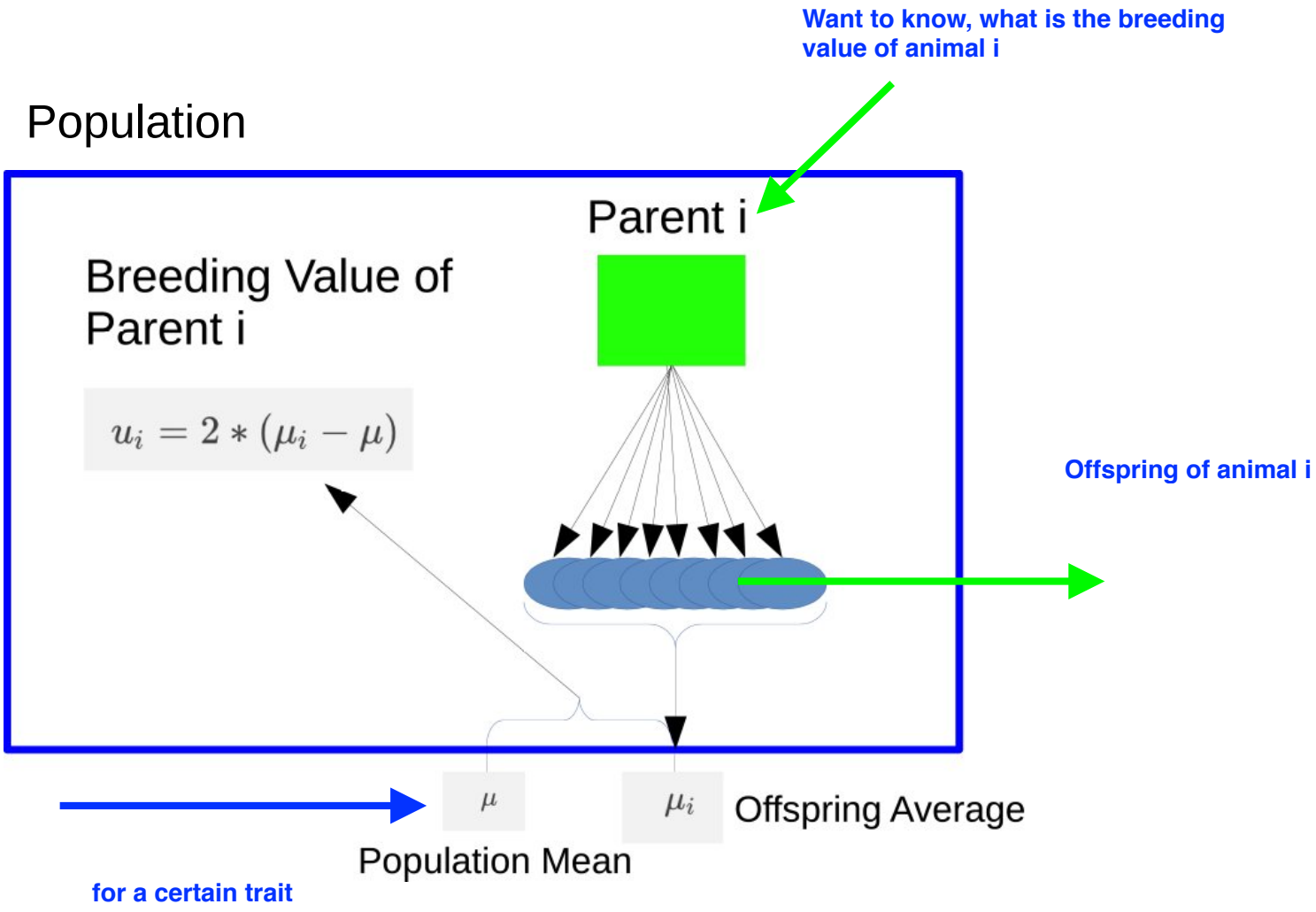
Peter von Rohr

04.05.2020

General Topic:
*** Assume that Breeding Organisation wants to include a new trait into the aggregate genotype**
*** What do we have to do?**
1. Model selection ==> results: important fixed effects and covariables to be used in the statistical model
2. Variance components estimation ==> results: estimates of variance components for all random effects
3. Predict breeding values ==> results: predicted breeding values are used as selection criterium

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

Definition is useful for verifcation of predicted breeding values

==> Top-bottom comparison

For two groups of sires
1. Sires with many offspring with a high predicted breeding value
2. Sires with many offspring with low predicted breeding value

Compare phenotypic observation of offspring of sires of the two groups
==> result: mean phenotypic values of offspring from the two different sire groups must be different

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.

represents the known environmental factors. In a more realistic model, this is replaced by a number of fixed effects.

Solution II

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

random effects

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

fixed effect

- Expected values and variance-covariance matrix

$$E\begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} = \begin{bmatrix} \mu \\ 0 \\ 0 \end{bmatrix}$$

Expected values of random effects are 0 because u_i and e_i are defined as deviations from the general mean

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

Diagonal elements are the variances of the random effects,
var(y_i) = \sigma_y^2
var(u_i) = \sigma_u^2
var(e_i) = \sigma_e^2

Offdiagonal elements are covariances: cov(y_i, u_i) = \sigma_u^2, cov(y_i, e_i) = \sigma_e^2, cov(u_i, e_i) = 0

How to Predict Breeding Values

Using a statistical method

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

Other examples: Least Squares Solution to fixed effects in a fixed model

\hat{b} = (X^TX)^{-1}X^Ty

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

function of the data y: Matrix (X^TX)^{-1}X^T multiplied with data vector 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

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

function of the data

- $\hat{u}$ consists of two parts

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

depends on population parameters such as the heritabilty and the genetic correlation between different traits or different breeding values

Is this useful as a predictor ==> looking at the properties

Unbiasedness

- Expected value ($E(\hat{u})$)

from the definition of predicted breeding value

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

constant factor

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

E(c * y) = c * E(y)

E(y - c) = E(y) - c

Variance

Goal: Prediction of a random effect u, with E(u) = 0 and var(u)
Expected value of the prediction is ok with unbiasedness
What about the variance of predicted breeding value
1. Want: variance of the predicted breeding value is as close as possible to var(u)
2: Want: covariance between true and predicted breeding value should as high as possible

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

Rule: var(A * y) = A * var(y) * A^T

Rule 2: var(y - c) = var(y)

$$\boxed{var(\hat{u})} = var(cov(u, y^T) * var(y)^{-1} * (y - E(y)))$$
$$= cov(u, y^T) * var(y)^{-1} * var(y - E(y))$$
$$* var(y)^{-1} * cov(y, u^T)$$
$$= cov(u, y^T) * var(y)^{-1} * cov(y, u^T)$$

=

Definition of predicted breeding value

$$\boxed{cov(u, \hat{u})} = cov(u, (cov(u, y^T) * var(y)^{-1} * (y - E(y)))^T)$$
$$= cov(u, (y - E(y))^T) * var(y)^{-1} * cov(y, u^T)$$
$$= cov(u, y^T) * var(y)^{-1} * cov(y, u^T) = var(\hat{u})$$

Rule 3: cov(u, A * y) = cov(u, y) * A^T

Rule 4: cov(u, (y-c)) = cov(u, y)

Accuracy

Correlation between true and predicted breeding value

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

Definition of a correlation

$$r_{u,\hat{u}} = \frac{cov(u, \hat{u})}{\sqrt{var(u) * var(\hat{u})}}$$

$$= \sqrt{\frac{var(\hat{u})}{var(u)}}$$

Want: variance of the predicted breeding value is as close as possible to var(u)
==> r is high

- Reliability ("Bestimmtheitsmass"): $B = r^2_{u,\hat{u}}$

B is used in practical livestock breeding as a measure of quality of each predicted breeding value. B is in most cases given by a percentage value

Prediction Error Variance (PEV)

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

Rule: var(a - b) = var(a) + var(b) - 2*cov(a,b)

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

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

Conditional Density

Example: Given two bulls S and T with the same predicted breeding value but with different reliabilities, Bull S has B = 99% and Bull T has B = 65% but Bull T is much younger than S, what is the risk of having low performing offspring when using Bull T instead of S

- 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 $var(u|\hat{u})$

$$\boxed{E(u|\hat{u})} = E(u) + cov(u, \hat{u}^T) * var(\hat{u})^{-1} * (\hat{u} - E(\hat{u})) \boxed{= \hat{u}}$$

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

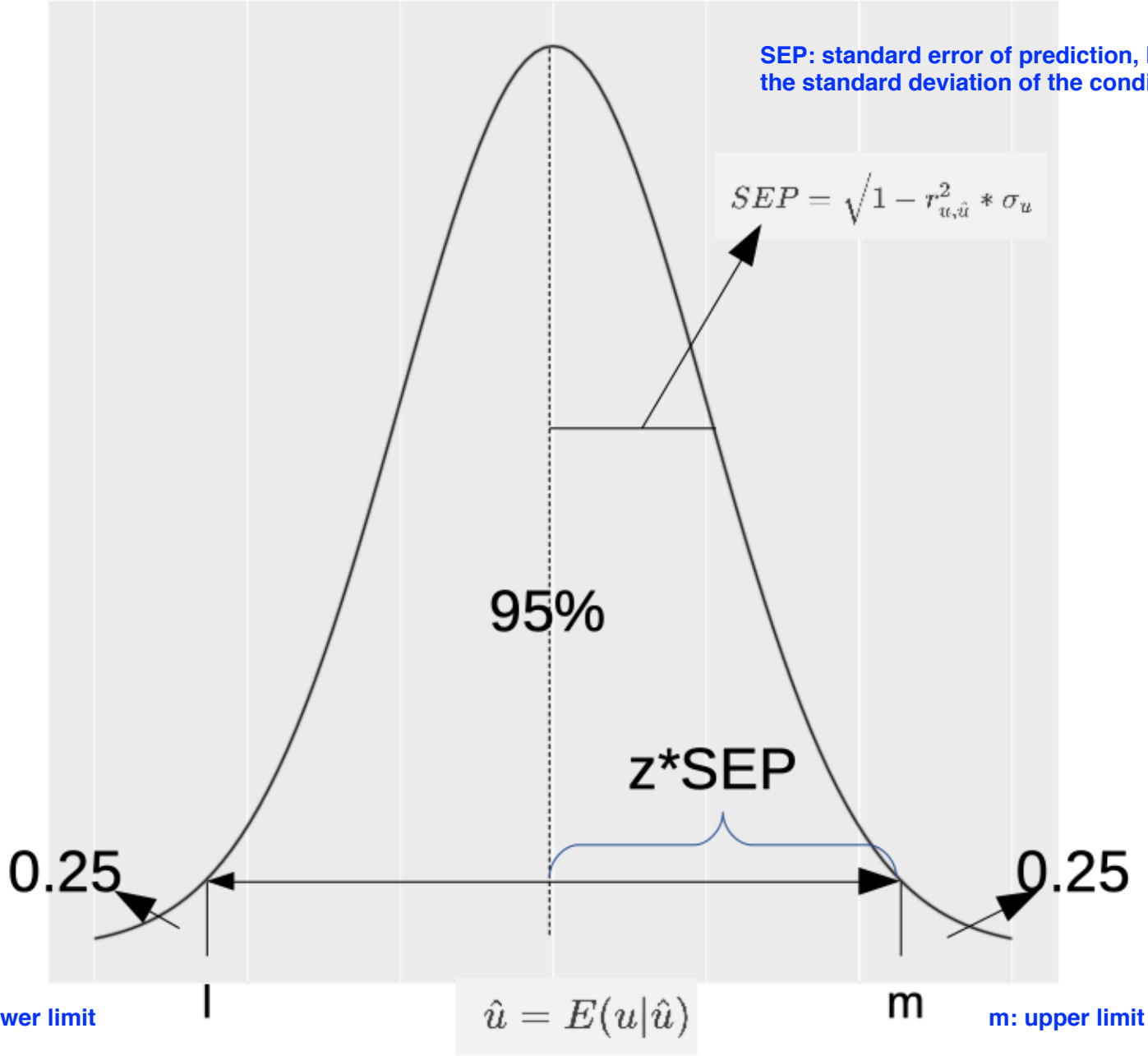
= PEV

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

Normal distribution represents the conditional density of f(u | \hat{u})



l: lower limit

m: upper limit of CI

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

For the 95% - CI, the value of z = 1.96

Linear Mixed Effects Model

New: Replace the general mean \mu by a set of fixed effects (b)

- 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

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

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

Solutions

- Same as for simple model

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

Correction of the y is now done with least squares solution of the fixed effects

with

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

corresponding to the general least squares solution of b

Problem

V = var(y) = ZGZ^T + R

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

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

R and its inverse have a simple structure, i.e. diagonal matrices
G and its inverse depend on the numerator relationship matrix and can be constructed directly without matrix inversion

Sire Model

$$y = Xb + Zs + e$$

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

$$var(s) = A_s * \sigma_s^2 \quad \mathbf{G}$$

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

$$var(a) = A\sigma_a^2 \quad \textbf{= G}$$

where A is the numerator relationship matrix