Variance Components Estimation

Peter von Rohr

27.04.2020

Genetic Variation

- Requirement for trait to be considered in breeding goal
- Breeding means improvement of next generation via selection and mating
- Only genetic (additive) components are passed to offspring
- Selection should be based on genetic component of trait
- Selection only possible with genetic variation

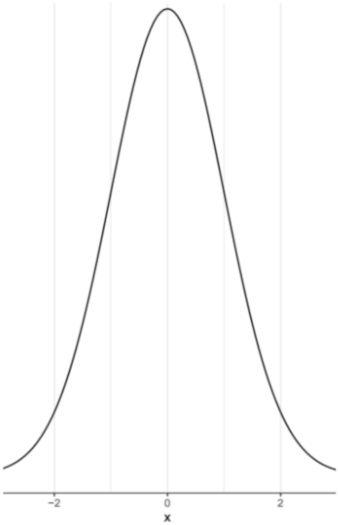
$\rightarrow$ genetic variation indicates how good characteristics are passed from parents to offspring

$\rightarrow$ measured by **heritability** $h^2 = \frac{\sigma_a^2}{\sigma_p^2}$

Two Traits

no variation

with variation



Problems

- Genetic components cannot be observed or measured
- Must be estimated from data
- Data are mostly phenotypic

$\rightarrow$ topic of variance components estimation

- Model based, that means connection between phenotypic measure and genetic component are based on certain model

$$p = g + e$$

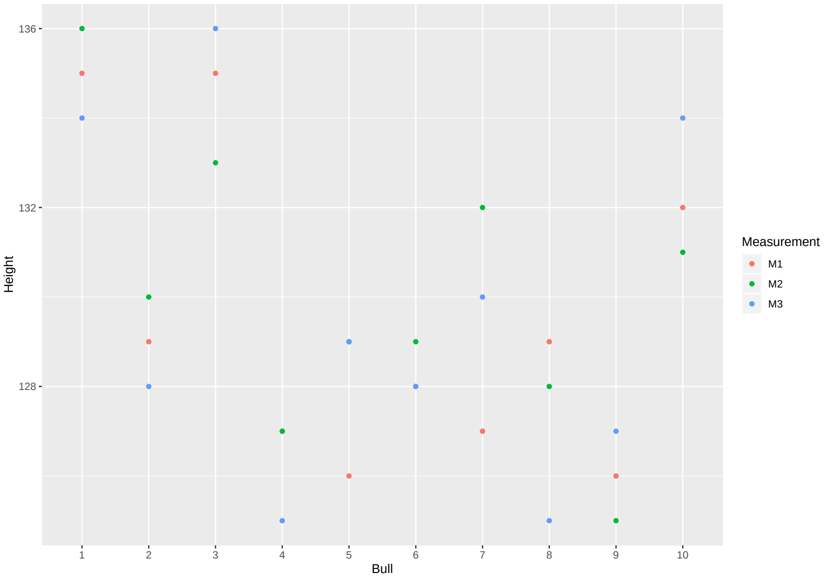
with $cov(g, e) = 0$

- **Goal**: separate variation due to g (σ_a^2) from phenotypic variation

Example of Variance Components Separation

- Estimation of repeatability
- Given repeated measurements of same trait at the same animal
- Repeatability means variation of measurements at the same animal is smaller than variation between measurements at different animals

Repeatability Plot



Model

$$y_{ij} = \mu + t_i + \epsilon_{ij}$$

where

y_{ij}	measurement j of animal i
μ	expected value of y
t_i	deviation of y_{ij} from μ attributed to animal i
ϵ_{ij}	measurement error

Estimation Of Variance Components

- $E(t_i) = 0$
- $\sigma_t^2 = E(t_i^2)$: variance component of total variance (σ_y^2) which can be attributed to the t-effects
- $E(\epsilon_{ij}) = 0$
- $\sigma_\epsilon^2 = E(\epsilon_{ij}^2)$: variance component attributed to ϵ-effects
- $\sigma_y^2 = \sigma_t^2 + \sigma_\epsilon^2$
- Repeatability w defined as:

$$w = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\epsilon^2}$$

$\rightarrow$ estimate of σ_t^2 needed

Analysis Of Variance (ANOVA)

Effect	df	Sum Sq	Mean Sq	$E(Mean\ Sq)$
Bull (t)	$r-1$	$SSQ(t)$	$SSQ(t)/(r-1)$	$\sigma_\epsilon^2 + n * \sigma_t^2$
Residual (ϵ)	$N-r$	$SSQ(\epsilon)$	$SSQ(\epsilon)/(N-r)$	σ_ϵ^2

where

$$SSQ(t) = \left[\frac{1}{n}\sum_{i=1}^{r}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right] - \left(\sum_{i=1}^{r}\sum_{j=1}^{n} y_{ij}\right)^2 / N$$

$$SSQ(\epsilon) = \sum_{i=1}^{r}\sum_{j=1}^{n} y_{ij}^2 - \left[\frac{1}{n}\sum_{i=1}^{r}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right]$$

Zahlenbeispiel

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Bull          9  286.7   31.85   13.85 8.74e-07 ***
## Residuals    20   46.0    2.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Setting expected values of `Mean Sq` equal to estimates of variance components

$$\hat{\sigma}_{\epsilon}^2 = 2.3 \text{ and } \hat{\sigma}_t^2 = \frac{31.85 - 2.3}{3} = 9.85$$

Repeatability

$$\hat{w} = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_{\epsilon}^2} = 0.81$$

Same Strategy for Sire Model

- Sire model is a mixed linear effects model with sire effects s as random components

$$y = Xb + Zs + e$$

- In case where sires are not related, $\$\mathrm{var(s)} = I * \sigma_s^2$
- From σ_s^2, we get genetic additive variance as $\sigma_a^2 = 4 * \sigma_s^2$

ANOVA

Effect	Degrees of Freedom	Sum Sq	Mean Sq	$E(Mean\ Sq)$
Sire $(s\|b)$	$r-1$	$SSQ(s\|b)$	$SSQ(s\|b)/(r-1)$	$\sigma_e^2 + k * \sigma_s^2$
Residual (e)	$N-r$	$SSQ(e)$	$SSQ(e)/(N-r)$	σ_e^2

with

$$k = \frac{1}{r-1}\left[N - \frac{\sum_{i=1}^{r} n_i^2}{N}\right]$$

Maximum Likelihood (ML)

- Likelihood

$$L(\theta) = f(y|\theta)$$

- Normal distribution

$$L(\theta) = (2\pi)^{-1/2n}\sigma^{-n}|H|^{-1/2} * exp\left\{-\frac{1}{2\sigma^2}(y - Xb)^T H^{-1}(y - Xb)\right\}$$

with $var(y) = H * \sigma^2$ and $\theta^T = \left[\begin{array}{cc} b & \sigma^2 \end{array}\right]$

Maximization of Likelihood

- Set $\lambda = logL$
- Compute partial derivatives of λ with respect to all unknowns

$$\frac{\partial \lambda}{\partial b}$$

$$\frac{\partial \lambda}{\partial \sigma^2}$$

- Set partial derivatives to 0 and solve for unknowns
- Use solutions as estimates

Restricted Maximum Likelihood (REML)

- Problem with ML: estimate of σ^2 depends on $b \to$ undesirable
- Do transformations Sy and Qy

(i) The matrix S has rank $n - t$ and the matrix Q has rank t

(ii) The result of the two transformations are independent, that means $cov(Sy, Qy) = 0$ which is met when $SHQ^T = 0$

(iii) The matrix S is chosen such that $E(Sy) = 0$ which means $SX = 0$

(iv) The matrix QX is of rank t, so that every linear function of the elements of Qy estimate a linear function of b.

REML II

- From (i) and (ii) it follows that the likelihood L of y is the product of the likelihoods of Sy (L^*) and Qy (L^{**}) that means

$$\lambda = \lambda^* + \lambda^{**}$$

- Variance components are estimated from λ^* which will then be independent of b

Bayesian Estimation

- Proposed already in the 80's
- Full implementation only in 1993
- Requirements:
  - cheap computing and
  - good pseudo-random number generators
- Bayesian estimation is based on conditional posterior distribution of unknowns given the knowns
- Conditional posterior distribution is computed from prior distribution of unknowns times the likelihood

Model

- Univariate Gaussian linear mixed model

$$y = Xb + Zu + e$$

where

y	vector of observations (length n)
b	vector of fixed effects (length p)
u	vector of random breeding values (length q)
e	vector of random residuals (length n)
X	$n \times p$ design matrix linking fixed effects to observations
Z	$n \times q$ design matrix linking breeding values to observations

Likelihood

- Data generating distribution

$$y|b, u, \sigma_e^2 \sim \mathcal{N}(Xb + Zu, I * \sigma_e^2)$$

where I is a $n \times n$ identity matrix and σ_e^2 is the variance of the random residuals.

Priors

- Prior distributions must be specified for all unknowns
- Unknowns in our example are: b, u, σ_e^2 and σ_u^2
- Prior distribution for
  - b is flat, i.e. $p(b) \propto c$
  - u Normal distribution as $u|G, \sigma_u^2 \sim N(0, G * \sigma_u^2)$
  - σ_e^2 scaled inverse χ^2:
    $p(\sigma_e^2|\nu_e, s_e^2) \propto (\sigma_e^2)^{-\nu_e/2-1} exp(-\frac{1}{2}\nu_e s_e^2/\sigma_e^2)$
  - σ_u^2 : $p(\sigma_u^2|\nu_u, s_u^2) \propto (\sigma_u^2)^{-\nu_u/2-1} exp(-\frac{1}{2}\nu_u s_u^2/\sigma_u^2)$
- ν_e, ν_s, s_e^2 and s_u^2 are called hyper-parameters and must be determined

Additional Terms

- Let

$$\theta^T = (b^T, u^T) = (\theta_1, \theta_2, \ldots, \theta_N)$$

$$\theta_{-i} = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_N)$$

- Further, let

$$s^T = (s_u^2, s_e^2)$$

and

$$\nu^T = (\nu_u, \nu_e)$$

Joint Posterior Density

The joint posterior distribution can be written as

$$p(\theta, \sigma_u^2, \sigma_e^2 | y, s, \nu) \propto p(\theta) * p(\sigma_u^2 | \nu_u, s_u^2) * p(\sigma_e^2 | \nu_e, s_e^2) * p(y | \theta, \sigma_e^2)$$

Fully Conditional Posterior Densities of θ

- Density of every single unknown component when setting all other components as known

$$\theta_i | y, \theta_{-i}, \sigma_u^2, \sigma_e^2, s, \nu \sim \mathcal{N}(\tilde{\theta}_i, \tilde{v}_i)$$

where $\tilde{\theta}_i = (r_i - \sum_{j=1, j \neq i}^{N} w_{ij}\theta_j)/w_{ii}$ and $\tilde{v}_i = \sigma_e^2 / w_{ii}$.

- vector r is the vector of right-hand side of MME
- matrix W is the coefficient matrix of MME

Fully Conditional Posterior Densities of σ_e^2

- scaled inverted chi-square distribution for σ_e^2

$$\sigma_e^2 | y, \theta, \sigma_u^2, s, \nu \sim \tilde{\nu_e} \tilde{s_e}^2 \chi^{-2}_{\tilde{\nu_e}}$$

- Parameters of the above distribution are defined as

$$\tilde{\nu_e} = n + \nu_e$$

and

$$\tilde{s_e}^2 = \left[(y - Xb - Zu)^T (y - Xb - Zu) + \nu_e s_e^2 \right] / \tilde{\nu_e}$$

Fully Conditional Posterior Densities of σ_u^2

- scaled inverted chi-square distribution for σ_u^2

$$\sigma_u^2 | y, \theta, \sigma_e^2, s, \nu \sim \tilde{\nu_u} \tilde{s_u}^2 \chi^{-2}_{\tilde{\nu_u}}$$

- Parameters of the above distribution are defined as

$$\tilde{\nu_u} = q + \nu_u$$

and

$$\tilde{s_u}^2 = \left[u^T G^{-1} u + \nu_u s_u^2 \right] / \tilde{\nu_u}$$

Implementation

- Step 1: set starting values for θ, σ_e^2 and σ_u^2
- Step 2: draw random number for each component θ_i of θ from fully conditional distribution $\mathcal{N}(\tilde{\theta}_i, \tilde{v}_i)$
- Step 3: draw random number for σ_e^2 from $\tilde{\nu_e}\tilde{s_e}^2\chi^{-2}_{\tilde{\nu_e}}$
- Step 4: draw random number for σ_u^2 from $\tilde{\nu_u}\tilde{s_u}^2\chi^{-2}_{\tilde{\nu_u}}$
- Repeat steps 2-4 many times and store random numbers
- Step 5: compute means of random numbers to get Bayesian estimates of unknowns θ, σ_e^2 and σ_u^2