Variance Components Estimation

Peter von Rohr

27.04.2020

Linear Mixed Effect model
> model selection ==> determination of fixed effects in the model
> variance components ==> random effect (breeding values and error)
> variance components estimation tells us how phenotypic variation is partiotioned among the random effects

Genetic Variation

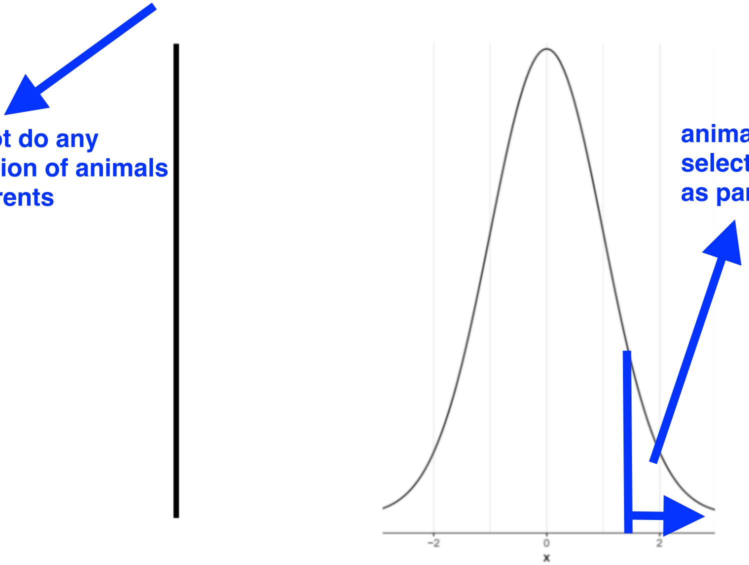
New trait: calf survival, mastitis resitence
Observe: phenotypic observation
In order to use a new trait in the breeding goal
the variation of the phenotypic observations must
in part be due to a genetic component

- 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

→ genetic variation indicates how good characteristics are passed from parents to offspring

→ measured by $\boxed{\textbf{heritability}\ h^2 = \frac{\sigma_a^2}{\sigma_p^2}}$

Two Traits



no variation

with variation

cannot do any selection of animals as parents

animals selected as parents

Problems

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

→ 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

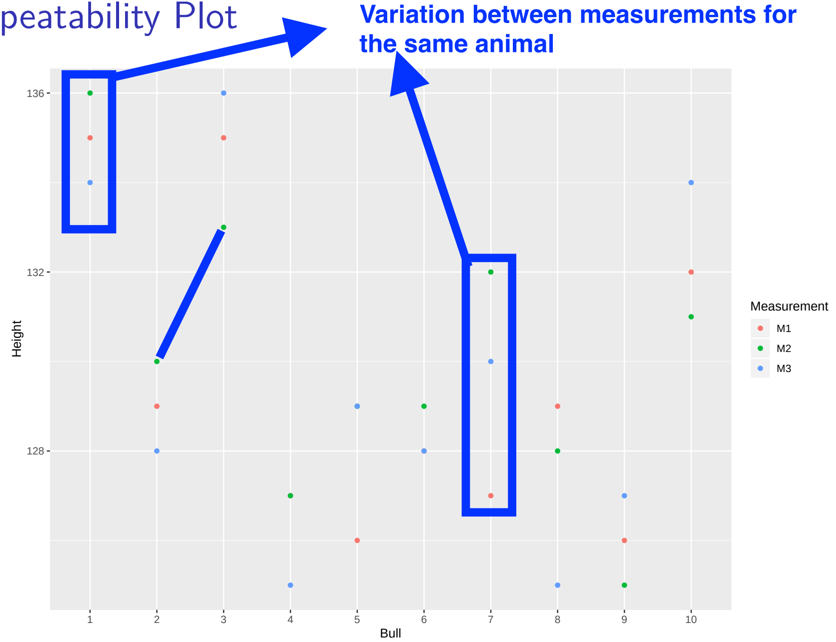
Example of Variance Components Separation

```
Simple dataset with repeated measurements of the same trait

Animal | Measurement 1 | Measurement 2 | Measurement 3
==================================================
1      |  y_{11}       | y_{12}        | y_{13}
2
3
```

- 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



Variation between measurements for the same animal

Model

phenotypic observation y_{ij}: height of the bull

effect that is the same for a given bull i

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

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

because t_i was defined as deviation from \mu

- $E(t_i) = 0$
- $\sigma_t^2 = E(t_i^2)$: 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
- $\sigma_y^2 = \sigma_t^2 + \sigma_\epsilon^2$ i.e., there are not covariances between t and errors
- Repeatability w defined as:

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

if w is high, then repeated measurements of the same trait do not give us additional information

→ estimate of σ_t^2 needed

Analysis Of Variance (ANOVA)

ANOVA - table

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

for each effect in the model, there is one row in the anova-table

SSQ: Sum of squares

(sum of all observation)^2 with N = total number of observations

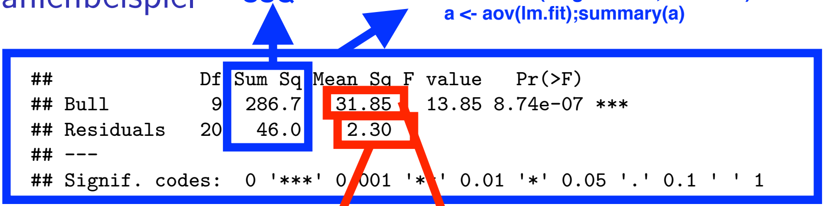
$$SSQ(t) = \left[\frac{1}{n}\sum_{i=1}^{r}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right] - \left(\sum_{i=1}^{r}\sum_{j=1}^{n} y_{ij}\right)^2 / N$$

sum of heights for bull i

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

sum of squared observations where r is the number of bulls and n is the number of measurements per bull

Zahlenbeispiel



ANOVA-Table obtained by R, aov()
lm.fit <- lm(height ~ bull, data = ...)
a <- aov(lm.fit);summary(a)

```
##             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

Computed values for the Mean Sq are set equal to the estimates of the variance components ==> \hat{\sigma_{\epsilon}^2 = Mean Sq(error)

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

Repeatability

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

Same Strategy for Sire Model

Principle of sire model
*** Observations are from daughters of different sires**
*** Comparable to the previous data set**

- 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, \$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

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

r: number of sires
n_i: the number of daughters for sire i
N: total number of observations

Problem: estimates of sire variance might get negative, because it is computed as a difference between MeanSQ(sire) - MeanSq(Residual) / …

Maximum Likelihood (ML)

We are forced to make an assumption of the distribution of the data.

- Likelihood

is defined as the conditional density of the the data (y) given the model parameters (\theta)

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

- Normal distribution

Assumed model: y = Xb + e

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

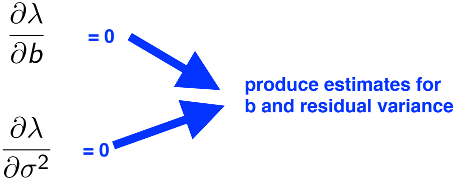
with $var(y) = H * \sigma^2$ and $\theta^T = \left[\begin{array}{cc} b & \sigma^2 \end{array}\right]$

Goal: Estimate the residual variance

Maximization of Likelihood

Principle of ML: Find the values of the parameters (b, residual variance) that maximize the value of the Likelihood function. The values for the unknown parameters that maximize the likelihood function are then used as Maxium-Likelihood estimates.

- $\boxed{\text{Set } \lambda = logL}$ → **to make computations easier**
- Compute partial derivatives of λ with respect to all unknowns



$$\frac{\partial \lambda}{\partial b} = 0$$

$$\frac{\partial \lambda}{\partial \sigma^2} = 0$$

produce estimates for b and residual variance

- 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 $b \rightarrow$ 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.

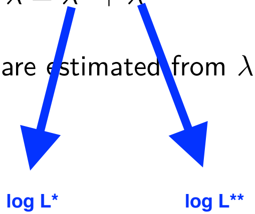
REML II

- 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

L* is the likelihood based on Sy
L is the likelihood based on Qy**

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

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



Bayesian Estimation

The estimation procedures so far (ANOVA, ML and REML) were all frequentist approaches

- 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: $\boxed{b,\ u,\ \sigma_e^2 \text{ and } \sigma_u^2}$
- 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

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

vector \theta except the i-th component

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

- 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

Proportional to ...

$$p(\theta, \sigma_u^2, \sigma_e^2 | y, s, \nu) \propto \underbrace{p(\theta) * p(\sigma_u^2 | \nu_u, s_u^2) * p(\sigma_e^2 | \nu_e, s_e^2)}_{\text{prior distributions that were specified for all the unknowns}} * \underbrace{p(y | \theta, \sigma_e^2)}_{\text{likelihood}}$$

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

for fixed effects and breeding values intial values are set to = 0

variance components can be initalised by half of the phenotypic variance

- 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 $\mathcal{N}(\tilde{\theta}_i, \tilde{v}_i)$ **in R : rnorm(1, mean = , sd =)**
- 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

Practical consideration
*** despite their ease of use, they are not commonly used**
*** reason: there is no reliable software around**