Variance Components Estimation

Peter von Rohr

04 December 2020

> Including a new trait in a breeding program always starts with estimation of variance components
> Goal: split the observable variation in phenotypic observations into their source according to the model that we want to use in our evaluations
> Need to determine a model:
 * fixed linear effect models: sources of variation: random residuals
 * mixed linear effect models: sources of variation: residuals, breeding values => genetic additive variance \sigma_u^2

Why

- Predictions of breeding values using BLUP requires variance components σ_u^2 or σ_s^2 and σ_e^2
- So far we have assumed that they are known
- In reality: must be estimated from data

Sire Model

- Start with a simple sire model

$$y = X\beta + Z_s s + e$$

with $var(e) = R$, $var(s) = A_s\sigma_s^2$ and $var(y) = Z_s A_s Z_s^T \sigma_s^2 + R$

- A_s: numerator relationship for sires
- σ_s^2 corresponds to $0.25 * \sigma_u^2$
- $R = I * \sigma_e^2$

$\rightarrow$ estimate σ_s^2 and σ_e^2 from data

Analysis of Variance (ANOVA)

Principle: Decomposition of sum of squares into the different source, and source are determined by the model

Sources of variation are all components that are present in a model

Why sum of squares? Fixed linear effect model: residual variance estimate is based on the sum of the square residuals.

Source	Degrees of Freedom (df)	Sums of Squares (SSQ)
Overall (μ)	$Rank(X) = 1$	$y^T X(X^T X)^{-1} X^T y = F$
Sires (s)	$Rank(Z_s) - Rank(X) = q - 1$	$y^T Z_s(Z_s^T Z_s)^{-1} Z_s^T y - y^T X(X^T X)^{-1} X^T y = S$
Residual (e)	$n - Rank(Z_s) = n - q$	$y^T y - y^T Z_s(Z_s^T Z_s)^{-1} Z_s^T y = R$
Total	n	$y^T y$

Sums of Squares

$$F = y^T X (X^T X)^{-1} X^T y = \frac{1}{n} \left[\sum_{i=1}^{n} y_i \right]^2$$

$$S = y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y - y^T X (X^T X)^{-1} X^T y = \sum_{i=1}^{q} \frac{1}{n_i} \left[\sum_{j=1}^{n_i} y_{ij} \right]^2 - F$$

$$R = y^T y - y^T Z_s (Z_s^T Z_s)^{-1} Z_s^T y = \sum_{i=1}^{n} y_i^2 - S - F$$

Estimates

Estimates of variance components \sigma_e^2 and \sigma_s^2 are obtained by replacing expected values of S and R by their observed values and by replacing the variance components by their estimates

- β and s fixed
- Estimates of σ_e^2 and σ_s^2 are based on observed sums of squares S and R
- Set their expected values equal to the observed sums of squares

$$E(R) = (n-q)\sigma_e^2$$

R = (n-q) * \hat{\sigma_e^2} ==> \hat{\sigma_e^2} = R /(n-q)

$$E(S) = (q-1)\sigma_e^2 + tr(Z_s M Z_s)\sigma_s^2$$

where $M = I - X(X^T X)^{-1} X^T$ and q is the number of sires.

$\rightarrow \widehat{\sigma_e^2} = \frac{R}{n-q}$ and $\widehat{\sigma_s^2} = \frac{S-(q-1)\widehat{\sigma_e^2}}{tr(Z_s M Z_s)}$

Numerical Example

Table 1: Small Example Dataset for Variance Components Estimation Using a Sire Model

Animal	Sire	WWG
4	2	2.9
5	1	4.0
6	3	3.5
7	2	3.5

- Model

$$y_{ij} = \mu + s_j + e_i$$

Design Matrices

$$X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \; Z_s = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

ANOVA

An analysis of variance can be constructed as

Source	Degrees of Freedom (df)	Sums of Squares (SSQ)
Overall (μ)	$Rank(X) = 1$	$F = 48.3025$
Sires (s)	$Rank(Z_s) - Rank(X) = q - 1$	$S = 0.4275$
Residual (e)	$n - Rank(Z_s) = n - q$	$R = 0.18$

Estimates

$$M = \begin{bmatrix} 0.75 & -0.25 & -0.25 & -0.25 \\ -0.25 & 0.75 & -0.25 & -0.25 \\ -0.25 & -0.25 & 0.75 & -0.25 \\ -0.25 & -0.25 & -0.25 & 0.75 \end{bmatrix}$$

$$Z_s^T M Z_s = \begin{bmatrix} 0.75 & -0.5 & -0.25 \\ -0.5 & 1 & -0.5 \\ -0.25 & -0.5 & 0.75 \end{bmatrix}$$

Results

$$\hat{\sigma_e^2} = R = 0.18$$

$$\hat{\sigma_s^2} = \frac{S - (q-1)\hat{\sigma_e^2}}{tr(Z_s^T M Z_s)} = \frac{0.4275 - 2 * 0.18}{2.5} = 0.027$$

Anova in R

- Assume dataset is stored in dataframe called `tbl_num_ex_chp12`

```
tbl_num_ex_chp12$Sire <- as.factor(tbl_num_ex_chp12$Sire)
aov_result <- aov(WWG ~ Sire, data = tbl_num_ex_chp12)
summary(aov_result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Sire         2 0.4275  0.2137   1.187  0.544
## Residuals    1 0.1800  0.1800
```

Problem with ANOVA: In certain datasets, estimates of variance components can get negative and they are therefore not valid, because variance components must be non-negative.

Likelihood

Conditional density of the observations
y given the parameter \mu and \Sigma.
Very often the density f is taken to be
a normal distribution, then \mu is the mean
and \Sigma is the variance.

- Definition of likelihood

$$L(\mu, \Sigma) = f(y|\mu, \Sigma)$$

with

$$f_Y(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} exp\left\{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right\}$$

multivariate normal distribution

Central Limit Theorem: the distribution of the sum of very many very small effects
will converge to a normal distribution.
Infinitesimal model: genetic additive effects (small, many)

Maximum Likelihood

- Maximize $L(\mu, \Sigma)$ with respect to Σ

$$\hat{\Sigma} = argmax_{\Sigma} L(\mu, \Sigma)$$

Bayesian Approach

- Estimates of unknown quantity Σ based on posterior distribution of unknowns given knowns
- Using Bayes Theorem:

$$\begin{aligned} f(\Sigma|y) &= \frac{f(\Sigma, y)}{f(y)} \\ &= \frac{f(y|\Sigma)f(\Sigma)}{f(y)} \\ &\propto f(y|\Sigma)f(\Sigma) \end{aligned}$$

where $f(\Sigma)$: prior distribution and $f(y|\Sigma)$: likelihood

Bayesian Estimates

- Fixed Linear Model with $\Sigma = \left[\begin{array}{c} \sigma_s^2 \\ \sigma_e^2 \end{array}\right]$
- Full conditional distributions
  - sire variance: $f(\sigma_s^2|\sigma_e^2, y)$ has a given standard distribution
  - residual variance: $f(\sigma_e^2|\sigma_s^2, y)$ has a given standard distribution
- Draw random numbers from full conditional distributions in turn
- Result will be samples from posterior distribution
- Estimates are computed as empirical means and standard deviation based on the samples, e.g for σ_s^2

$$\widehat{\sigma_s^2}_{Bayes} = \frac{1}{N}\sum_{t=1}^{N}(\sigma_s^2)^{(t)}$$