Variance component estimation

Winter School 2022

Jian Zeng

Outline

1. Reminder Linear models
2. Balanced designs; one-way models; the sire model
3. Balanced designs; one-way models; ANOVA
   - ANOVA: ANalysis Of VAriance
   - Properties of estimators
4. Maximum likelihood approaches - ML and REML
   - Why can't we focus only on ANOVA approaches?
   - Maximum likelihood
   - Residual maximum likelihood (REML)
5. ML vs REML
6. Link to Heritability

Outline

Fixed models

Assume we have N groups of T individuals each,

$$y_{it} = \mu + \beta_i + e_{it}, \qquad i = 1, \ldots, N \text{ and } t = 1, \ldots, T$$

where

- μ is an unknown parameter to estimate (fixed effect)
- β_i is an unknown parameter to estimate that is constant for all t at i fixed (fixed effects)
- e_{it} residuals, with mean $E(\boldsymbol{e}) = 0$ and variance-covariance $\mathrm{Var}(\boldsymbol{e}) = \sigma_e^2 \boldsymbol{I}$. $\mathrm{Var}(y_{it}) = \mathrm{Var}(e_{it}) = \sigma_e^2$

Suppose we have 2 groups ($i = 1, 2$) of 3 individual ($t = 1, 2, 3$).

$$y_{it} = \mu + \beta_i + e_{it}, \text{ or in matrix form } \boldsymbol{y} = \mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{\beta}$ is a vector of fixed effects. The levels represents all the levels of interest,

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

Random models

Assume we have N groups of T individuals each,

$$y_{it} = \mu + u_i + e_{it}, \qquad i = 1, \ldots, N \text{ and } t = 1, \ldots, T$$

where

- μ is an unknown parameter to estimate (fixed effect)
- $(u_1, \ldots, u_N)$ is a vector or random values with mean $E(\boldsymbol{u}) = 0$ and variance-covariance $\mathrm{Var}(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I}$ (random effect)
- e_{it} residuals, with mean $E(\boldsymbol{e}) = 0$ and variance-covariance $\mathrm{Var}(\boldsymbol{e}) = \sigma_e^2 \boldsymbol{I}$. $\mathrm{Var}(y_{it}) = \mathrm{Var}(u_i) + \mathrm{Var}(e_{it}) = \sigma_u^2 + \sigma_e^2$

Suppose we have 2 groups ($i = 1, 2$) of 3 individual ($t = 1, 2, 3$).

$$y_{it} = \mu + u_i + e_{it} \text{ or in a matrix form } \boldsymbol{y} = \mu + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{u}$ is random, the levels are considered drawn from an infinite population of levels,

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu \\ \mu \\ \mu \\ \mu \\ \mu \\ \mu \end{pmatrix}, \boldsymbol{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

Mixed models

Mixed models (MM) contain both fixed and random factors

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

where

- $\boldsymbol{y}$ vector of observed dependent values, with mean $E(\boldsymbol{y}) = \boldsymbol{X}\beta$
- β vector of unknown parameters to estimate (fixed effects)
- $\boldsymbol{u}$ vector of unknown random effects, with mean $E(\boldsymbol{u}) = 0$ and variance-covariance $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{G}$ (usually $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}$)
- $\boldsymbol{e}$ vector of residuals, with mean $E(\boldsymbol{e}) = 0$ and variance-covariance $\mathrm{Var}(\boldsymbol{e}) = \boldsymbol{R}$ (usually $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}$),
- $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices

$$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R}$$

Example

Fitting a mean, unrelated sires, uncorrelated errors

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Zu} + \boldsymbol{e}$$

$$\boldsymbol{V} = \mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{Z}^\top \sigma_u^2 + \boldsymbol{I}\sigma_e^2$$

If $\boldsymbol{Z} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$, then $\boldsymbol{Z}\boldsymbol{Z}^\top = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$

And we get $\boldsymbol{V} = \begin{pmatrix} \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \sigma_u^2 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & 0 & 0 & 0 \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_u^2 + \sigma_e^2 & \sigma_u^2 & \sigma_u^2 \\ 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 & \sigma_u^2 \\ 0 & 0 & 0 & \sigma_u^2 & \sigma_u^2 & \sigma_u^2 + \sigma_e^2 \end{pmatrix}$

Aim

GBLUP reminder

$$\boldsymbol{y} = 1_n\mu + \boldsymbol{Zg} + \boldsymbol{e}$$

breeding values $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{G}\sigma_a^2)$, where $\boldsymbol{G}$ is the genomic relationship matrix.
GBLUP solves the following system of equations

$$\begin{bmatrix} 1_n'1_n & 1_n'\boldsymbol{Z} \\ \boldsymbol{Z}'1_n & \boldsymbol{Z}'\boldsymbol{Z} + \boldsymbol{G}^{-1}\frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} 1_n'\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{y} \end{bmatrix}$$

BLUP: estimate the mean μ and predict breeding values $\boldsymbol{g}$, all based on known variance

Aim of these lectures: estimate variance components (estimate λ)

Two ways to estimate variance components:

- ANOVA: ANalysis Of VAriance
- Maximum Likelihood approaches

Outline

Sire model

Consider a sire model with s unrelated sires and n progeny per sire; one-way model (one random effect: sire), with a balanced design (same number of observations per sire).

$y_{it} = \mu + u_i + e_{it}$
$\mathrm{Var}(y_{it}) = \sigma_u^2 + \sigma_e^2$
$\mathrm{Cov}(y_{it}, y_{ik}) = \sigma_u^2$
$i = 1 \ldots, s$
$t = 1, \ldots, n$

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$
$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{Z}^\top \sigma_u^2 + \boldsymbol{I}\sigma_e^2$
$\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i \boldsymbol{Z}_i^\top \sigma_u^2$
$\boldsymbol{X}$: matrix of one column of sn ones
$\boldsymbol{\beta} = \mu$

```
## Simulating data
> set.seed(123)
> mu = 1.2; s = 10; n = 20
> sigmau = 1; sigmae = 0.2
> ui = rnorm(s,sd=sigmau)
> eit = rnorm(n*s, sd=sigmae)
> y = matrix(NA_real_, nrow=n*s, ncol=1)
> for(i in 1:s){
+     ind = (n*(i-1)+1) : (n*(i-1)+n)
+     y[ind] = mu + ui[i] + eit[ind]
+ }

## create a grouping factor
> grp = factor(rep(1:s,each = n))

## create Z
> Z = matrix(0,nrow=s*n, ncol=s)
> for(i in 1:s){
+     ind = which(grp = = i)
+     Z[indic,i] = 1
+ }
```

Outline

Outline

ANOVA table - general framework

Source of variation	df	SS	MS	E(MS)
Mean	$df_M = 1$	SSM	SSM/df_M	$\mathrm{E}(SSM/df_M)$
Between sires	$df_B = s - 1$	SSB	SSB/df_S	$\mathrm{E}(SSB/df_S)$
Within sires (residuals)	$df_W = s(n-1)$	SSW	SSW/df_W	$\mathrm{E}(SSW/df_W)$
Total	N	SST		

df: degrees of freedom
SS: sum of squares
MS: Mean square (mean of SS)
E(MS): Expectation of MS

Sums of squares

We decompose SST into a mean, a between and a within family component:

$$\begin{aligned} SST &= \sum_i \sum_t y_{it}^2 = \sum_i \sum_t (y_{it} - \bar{y} + \bar{y})^2 \\ &= \sum_i \sum_t (y_{it} - \bar{y})^2 + N\bar{y}^2 \\ &= \sum_i \sum_t [(y_{it} - \bar{y}_i) + (\bar{y}_i - \bar{y})]^2 + N\bar{y}^2 \\ &= \sum_i \sum_t \left[(y_{it} - \bar{y}_i)^2 + 2(y_{it} - \bar{y}_i)(\bar{y}_i - \bar{y}) + (\bar{y}_i - \bar{y})^2\right] + N\bar{y}^2 \end{aligned}$$

$\bar{y}$ is the grand mean, $\bar{y}_i$ is the family mean.

The middle term is equal to zero by definition of a mean: $\sum_t (y_{it} - \bar{y}_i) = 0$.
The last term is independent of j so $\sum_i \sum_t (\bar{y}_i - \bar{y})^2 = n \sum_i (\bar{y}_i - \bar{y})^2$

Decomposition of SS

$$\begin{aligned} SST &= N\bar{y}^2 + n\sum_i (\bar{y}_i - \bar{y})^2 + \sum_i \sum_t (y_{it} - \bar{y}_i)^2 \\ &= SS_M + SS_B + SS_W \end{aligned}$$

ANOVA table - our sire model

Using

$$SSM = (\sum_i \sum_t y_{it})^2/(sn)$$

$$SSB = \sum_i (\sum_t y_{it})^2/n - (\sum_i \sum_t y_{it})^2/(sn)$$

$$SSW = \sum_i \sum_t y_{it}^2 - SSB - SSM$$

$$SST = \sum_i \sum_t y_{it}^2$$

we get

Source of variation	df	SS	MS	E(MS)
Mean	1	SSM	$SSM/1$	$N\mu^2 + n\sigma_u^2 + \sigma_e^2$
Between sires	$s-1$	SSB	$B = SSB/(s-1)$	$n\sigma_u^2 + \sigma_e^2$
Within sires (residuals)	$s(n-1)$	SSW	$W = SSW/(s(n-1))$	σ_e^2
Total	$N = sn$	SST		

cf lecture notes for calculations and proofs

Sum of Squares - Sire model

$$SST = \sum_i \sum_t y_{it}^2$$

$$SSM = (\sum_i \sum_t y_{it})^2/(sn)$$

$$SSB = \sum_i (\sum_t y_{it})^2/n - (\sum_i \sum_t y_{it})^2/(sn)$$

$$SSW = \sum_i \sum_t y_{it}^2 - SSB - SSM$$

```
> SST = sum(y^2); SST
[1] 422.0067

> SSM = sum(y)^2 / (s*n); SSM
[1] 279.1494

> SSB = 0
> for(i in 1:s){
+     ind = (n*(i-1)+1) : (n*(i-1)+n)
+     yit = sum(y[ind])
+     SSB = SSB + yit^2/n
+ }
> SSB = SSB - SSM; SSB
[1] 133.8501

> SSW = sum(y^2) - SSB - SSM; SSW
[1] 9.007134
```

Estimation of σ_u^2 and σ_e^2

Principle of ANOVA

"Equate SS of analysis of variance to their expected values, giving a set of equations that are linear in the variance components to be estimated"

For the one-way design: two equations, two unknowns:

$$\begin{aligned} B &= n\sigma_u^2 + \sigma_e^2 \\ W &= \sigma_e^2 \end{aligned}$$

Hence,

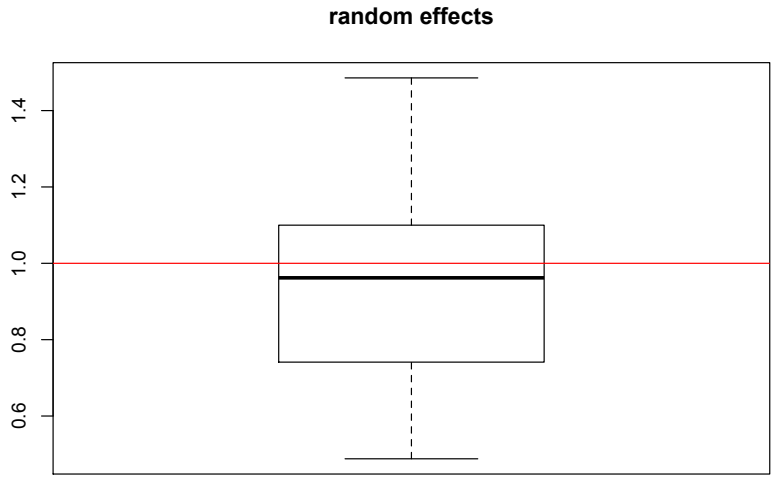
$$\begin{aligned} \widehat{\sigma_e^2} &= W \\ \widehat{\sigma_u^2} &= (B - W)/n \end{aligned}$$

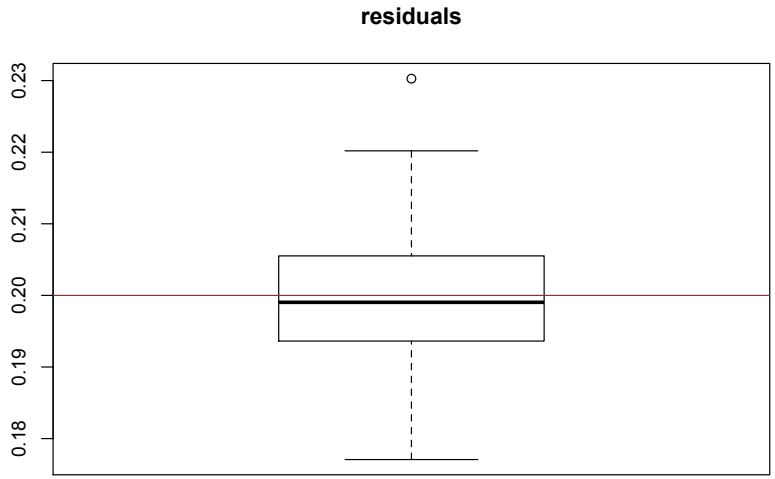
```
> B = SSB/(s-1); B
[1] 14.87224

> W = SSW/(s*(n-1)); W
[1] 0.04740597


> c(sqrt(W),sigmae)
[1] 0.2177291 0.2000000

> c(sqrt((B - W)/n),sigmau)
[1] 0.8609538 1.0000000
```

Outline

50 repeats - no(?) Bias

```
> u = e = NULL
> for(iter in 1:50){
+     ui = rnorm(s,sd=sigmau)
+     eit = rnorm(n*s, sd=sigmae)
+
+     y = mu + Z %* % ui + eit
+
+     SST = sum(y^2)
+     SSM = t(y) %*% X %*%
+         solve(t(X) %*% X) %*%
+         t(X) %*% y
+     SSB = t(y) %*% Z %*%
+         solve(t(Z)%*%Z) %*%
+         t(Z) %*% y - SSM
+     SSW = sum(y^2) - SSB - SSM
+
+     B = SSB/(s-1); B
+     W = SSW/(s*(n-1)); W
+
+     e = c(e, sqrt(W))
+     u = c(u, sqrt((B - W)/n))
+ }
```





Properties of estimators

1. Unbiased

$$\begin{aligned} E(\widehat{\sigma_e^2}) &= E(W) = \sigma_e^2 \\ E(\widehat{\sigma_u^2}) &= E(B-W)/n = E(B)/n - E(W)/n \\ &= (\sigma_u^2 + \sigma_e^2/n) - \sigma_e^2/n \\ &= \sigma_u^2 \end{aligned}$$

2. Minimum Variance: Estimates have minimum variance among all possible unbiased estimators. True for normal and non-normal data
3. Distribution: Under normality, only the estimate of the residual variance has a χ^2 distribution

4. Sampling variances

Using that $SSW \sim \sigma_e^2 \chi^2(df_W)$, and $\mathrm{Var}(SSW) = 2\sigma_e^4 df_W$, we have

$$\begin{aligned}
\mathrm{Var}(\widehat{\sigma_e^2}) &= \mathrm{Var}(W) = \mathrm{Var}(SSW)/df_w^2 \\
&= df_W 2\sigma_e^4/df_W^2 \\
&= 2\sigma_e^4/df_W \\
\mathrm{Var}(\widehat{\sigma_u^2}) &= \mathrm{Var}((B-W)/n) \\
&= [\mathrm{Var}(B) + \mathrm{Var}(W)]/n^2 \\
&= \left[2E(B)^2/df_B + 2E(W)^2/df_W\right]/n^2 \\
&= \left[2(n\sigma_u^2 + \sigma_e^2)^2/df_B + 2\sigma_e^4/df_W\right]/n^2 \\
&= (2/n^2)\left[(n\sigma_u^2 + \sigma_e^2)^2/(s-1) + \sigma_e^4/(s(n-1))\right]
\end{aligned}$$

$$\mathrm{Var}(\hat{t}) = \mathrm{Var}((B-W)/(B+(n-1)W))$$

```
> var(e^2)
[1] 1.728204e-05
> 2*sigmae^4/(s*(n-1))
[1] 1.684211e-05
>
> var(u^2)
[1] 0.2296663
> 2/n^2 * ( (n*sigmau^2+sigmae^2)^2/(s-1) + sigmae^4/(s*(n-1))   )
[1] 0.223112
```

Outline

Outline

Problems with Unbalanced Designs

- SS can be partitioned in many ways
  - fit effect α before β
  - fit effect β before α
- no obvious SS (or other quadratic form) to estimate variance components from
- SS (MS) are not *orthogonal* (independent)
- using an ANOVA approach gives biased estimates of variance components for a mixed model

Methods proposed for unbalanced designs

- Henderson's (1953) methods I, II, and III
  - Essentially Least Squares methods
  - Problems with mixed models
- Maximum Likelihood (ML)
  - Unified procedure for estimating fixed effects and variance components
  - Desirable asymptotic properties
  - Bias in variance components
- Residual (restricted) Maximum Likelihood (REML)
  - Similar to ANOVA for balanced designs
  - No bias due to loss in degrees of freedom for fitting fixed effects

Outline

Log-likelihood - general case

Our model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ can also be written as a generalised linear model (GLM)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \text{ where } \boldsymbol{e} \sim \mathcal{N}_n(0, \boldsymbol{V})$$

with $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R}$, where usually $\boldsymbol{G} = \sigma_u^2 \boldsymbol{I}$ and $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}$.

The likelihood of such model is

$$L(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2) = \left(\frac{1}{2\pi}\right)^{N/2} |\boldsymbol{V}|^{-1/2} \exp\left(\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$$

The log-likelihood in the general case is

$$\ell(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}\left[\log(|\boldsymbol{V}|) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]$$

The log-likelihood for the sire model is

$$\ell(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}\left[\log(|\boldsymbol{V}|) + (\boldsymbol{y} - 1\boldsymbol{\mu})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - 1\boldsymbol{\mu})\right],$$

and $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top + \boldsymbol{R} = \boldsymbol{Z}\boldsymbol{Z}^\top\sigma_u^2 + \boldsymbol{I}\sigma_e^2$.

Maximum Likelihood approach - one-way model

It follows (Searle, Linear Models, page 418) that,

$$\ell(\beta, \sigma_u^2, \sigma_e^2) = -\frac{1}{2}\left[s\log(n\sigma_u^2 + \sigma_e^2) + s(n-1)\log(\sigma_e^2) + SSW/\sigma_e^2 + SSB/(n\sigma_u^2 + \sigma_e^2) + sn(\bar{y} - \mu)^2/(n\sigma_u^2 + \sigma_e^2)\right]$$

Maximum Likelihood Estimation (MLE)

Taking differentials with respect to μ, σ_u^2 and σ_e^2, we obtain

$$\begin{aligned} \widehat{\mu} &= \bar{y} \\ \widehat{\sigma}_u^2 &= \left((\frac{s-1}{s})B - W\right)/n \\ \widehat{\sigma}_e^2 &= W \end{aligned}$$

with the condition that $((s-1)/s)B \geq W$

```
> c(sum(y)/(n*s), mu)
[1] 1.275278 1.200000

> c(((s-1)/s*B-W)/n, sigmau)
[1] 0.7988936 1.0000000

> c(sqrt(W), sigmae)
[1] 0.1923142 0.2000000
```

$=>$ classic estimate of the mean, same estimate for σ_e^2 as ANOVA, but biased estimate of σ_u^2 ((B-W)/n for anova)

Outline

ML vs REML

Maximum Likelihood (ML)

$$\ell = -\left[(\log(|\boldsymbol{V}|) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right]/2$$

'Determinant of a variance matrix, plus a weighted sum of squares of residuals'

Residual (or restricted) Maximum Likelihood (REML)

$$\ell_R = -\left[(\log(|\boldsymbol{V}|) - \log(|\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}|^{-1}) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\right]/2$$

$-\log(|\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}|^{-1})$: variance term associated with the estimation of $\widehat{\boldsymbol{\beta}}$
"penalty term"

Where does that come from?

Penalty term

Marginal model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

- $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$
- $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top\sigma_u^2 + \boldsymbol{R}\sigma_e^2$
- $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top\sigma_u^2 + \boldsymbol{R}\sigma_e^2)$

Linear combinations of $\boldsymbol{y}$ have a non-negative variance ($\boldsymbol{V}$ is 'non-negative definite')

If we knew the matrix $\boldsymbol{V}$, then
$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{y}$ (weighted least squares)
$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$

Side note: the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ is more restrictive than the marginal model: variances of both $\boldsymbol{e}$ and $\boldsymbol{u}$ are non-negative.

Estimation/computation

How to maximise the likelihood?

- Maximise likelihood = maximise log-likelihood
- Many methods, e.g.,
  - Derivative free
  - E(xpectation)M(aximisation)
  - Using second differentials

Fixed effects and REML

Usual estimates are,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{y}$$

These estimates of fixed effects are **not** maximum likelihood estimates!!!

- The likelihood was optimised independent of the fixed effects
- ML properties for the estimates of fixed effects do not apply $->$ no LRT, e.g. use Wald test

REML for balanced one-way model

$$\ell_R = -\frac{1}{2}\left[(s-1)\log(n\sigma_u^2+\sigma_e^2) + s(n-1)\log(\sigma_e^2) + SSW/\sigma_e^2 + SSB/(n\sigma_u^2+\sigma_e^2)\right]$$

We solve the partial derivatives $\frac{\partial \ell_R}{\partial \sigma_e^2}$ and $\frac{\partial \ell_R}{\partial \sigma_u^2}$ and obtain

REML estimates

$$\begin{aligned} \widehat{\sigma_e^2} &= W \\ \widehat{\sigma_u^2} &= (B-W)/n \end{aligned}$$

Same as ANOVA estimates (balanced, one way)!

Estimates

- If $B > W$, then the maximum likelihood estimates are identical to the ANOVA estimates
- If $B < W$ (negative ANOVA estimates), then

$$\begin{aligned}
\widehat{\sigma_u^2} &= 0 \\
\widehat{\sigma_e^2} &= (w_1 W + w_2 B)/(w_1 + w_2) \\
&= (SSB + SSW)/(ns - 1) \\
&= TSS/(N - 1)
\end{aligned}$$

where the weights w_i =1/variance.

$$\begin{aligned}
w_1 &= 1/(2\sigma_e^2/(s(n-1))) \\
w_2 &= 1/(2\sigma_e^4/(s-1))
\end{aligned}$$

Outline

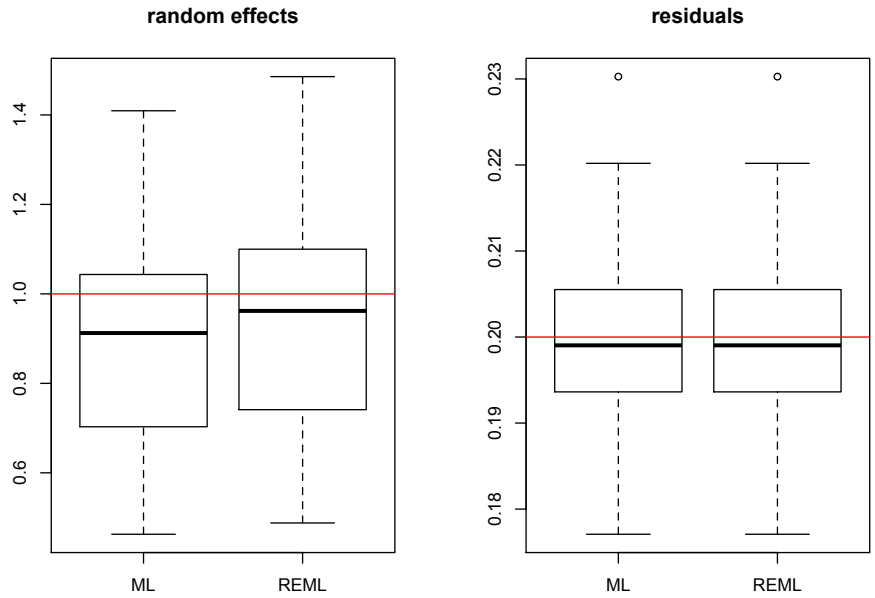
Balanced designs; on-way models

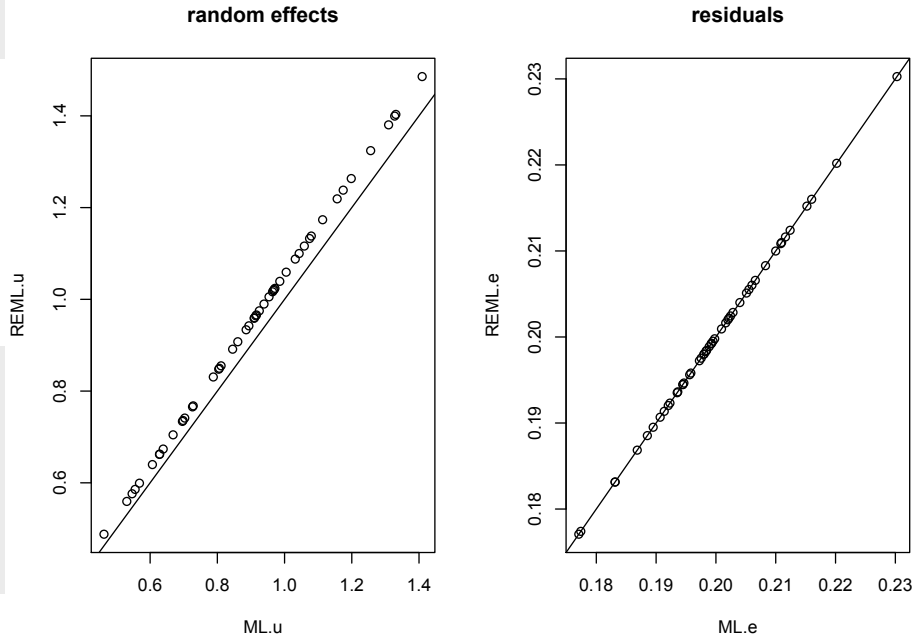
$$\begin{aligned}\widehat{\mu} &= \bar{y} \\ \widehat{\sigma^2_u} &= \left((\frac{s-1}{s})B - W \right) /n \\ \widehat{\sigma^2_e} &= W\end{aligned}$$

```
> c(sum(y)/(n*s), mu)
[1] 1.275278 1.200000

> c(((s-1)/s*B-W)/n, sigmau)
[1] 0.7988936 1.0000000

> c(sqrt(W), sigmae)
[1] 0.1923142 0.2000000
```

$$\begin{aligned}\widehat{\sigma^2_e} &= W \\ \widehat{\sigma^2_u} &= (B - W)/n\end{aligned}$$

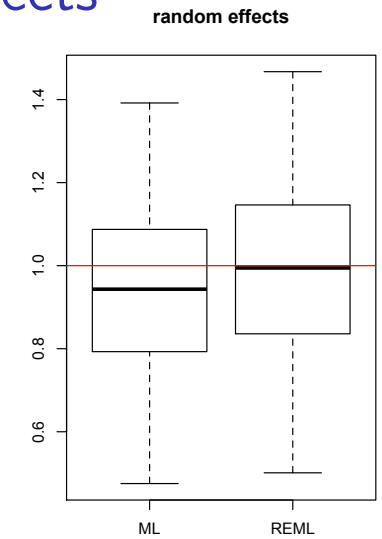
```
> c(sqrt(W),sigmae)
[1] 0.2177291 0.2000000

> c(sqrt((B - W)/n),sigmau)
[1] 0.8609538 1.0000000
```

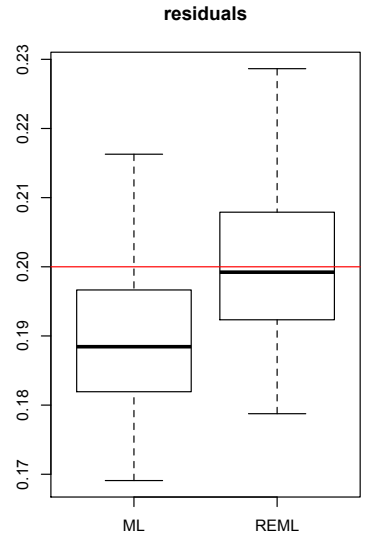


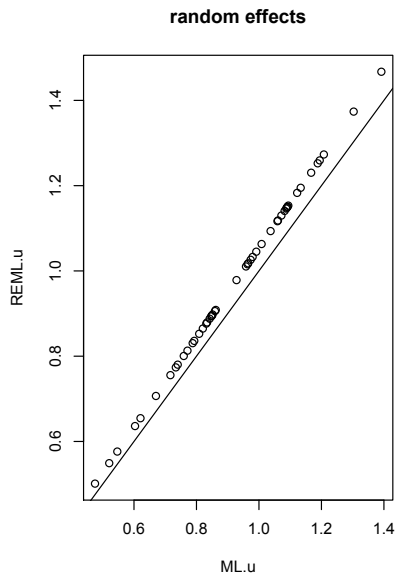


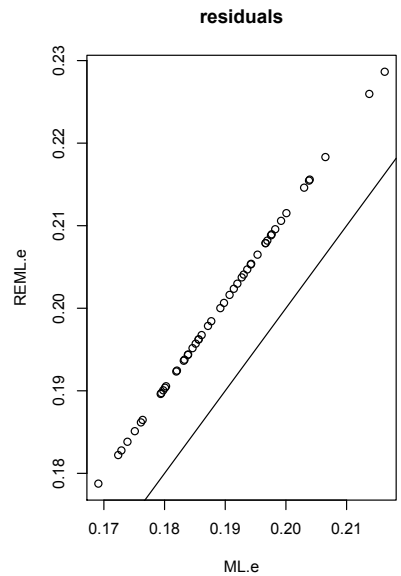
Balanced designs; on-way models; fixed effects

```
> library(lme4)
> p = 20
> X = matrix(rnorm(s*n*p), nrow=s*n)
> X = scale(X)
> beta = rnorm(p, sd=0.1)

> ML.u=ML.e=
+ REML.u=REML.e= vector("numeric", length=50)
> for(iter in 1:50){
+     ui = rnorm(s,sd=sigmau)
+     eit = rnorm(n*s, sd=sigmae)
+
+     y = mu + X%*% beta + Z %*% ui + eit
+
+     a = lmer(y~ X+ (1|grp), REML=FALSE) # ML
+     b = lmer(y~ X+ (1|grp), REML=TRUE) # REML
+
+     ML.u[iter] = as.data.frame(VarCorr(a))[1,5]
+     ML.e[iter] = as.data.frame(VarCorr(a))[2,5]
+     REML.u[iter] = as.data.frame(VarCorr(b))[1,5]
+     REML.e[iter] = as.data.frame(VarCorr(b))[2,5]
+ }
```



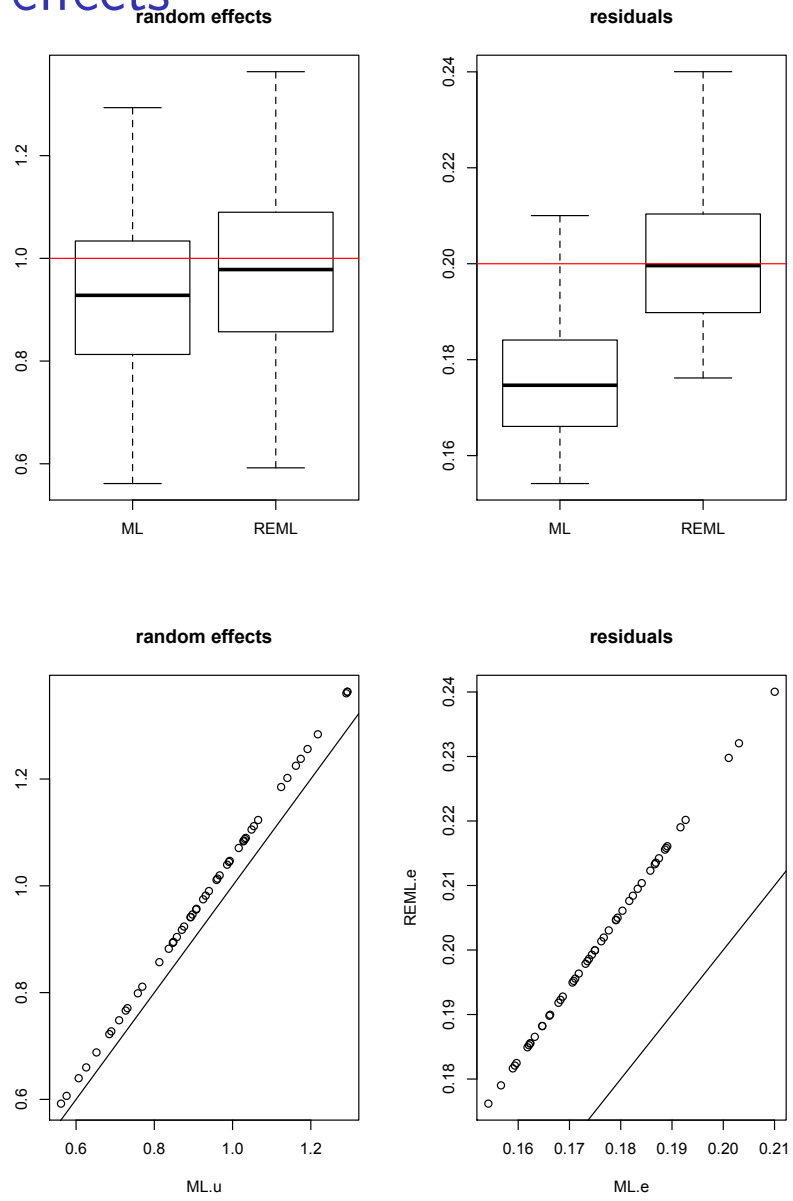






Unbalanced designs; one-way models; fixed effects

```
> n = sample(5:20,s, replace=TRUE)
> grp=factor(rep(1:s, n))

> ML.u=ML.e=
+ REML.u=REML.e= vector("numeric", length=50)
> for(iter in 1:50){
+     ui = rnorm(s,sd=sigmau)
+     eit = rnorm(n*s, sd=sigmae)
+
+     y = mu + X%*% beta + Z %*% ui + eit
+
+     a = lmer(y~ X+ (1|grp), REML=FALSE) # ML
+     b = lmer(y~ X+ (1|grp), REML=TRUE) # REML
+
+     ML.u[iter] = as.data.frame(VarCorr(a))[1,5]
+     ML.e[iter] = as.data.frame(VarCorr(a))[2,5]
+     REML.u[iter] = as.data.frame(VarCorr(b))[1,5]
+     REML.e[iter] = as.data.frame(VarCorr(b))[2,5]
+ }
```



Outline

Genome-Wide Complex Trait Analysis (GCTA)

$$\boldsymbol{y} = 1_{\boldsymbol{n}}\mu + \boldsymbol{g} + \boldsymbol{e}, \tag{1}$$

$\mathrm{Var}(\boldsymbol{g}) = \boldsymbol{G}\sigma_a^2$ with $\boldsymbol{G}$ the relatedness (GRM) matrix; $\boldsymbol{V} = \mathrm{Var}(\boldsymbol{y}) = \boldsymbol{G}\sigma_g^2 + \boldsymbol{I}\sigma_e^2$.

σ_g^2 and σ_e^2 estimated by REML; $h^2 = \dfrac{\sigma_g^2}{\mathrm{Var}(\boldsymbol{y})}$.

limitation of R: `lmer`: random effects are assumed independent; you cannot input the GRM. `lmer4qtl` seems to be a recent alternative.
In any case, R is not advised for ML/REML analysis with big datasets (too slow)

We use `GCTA`