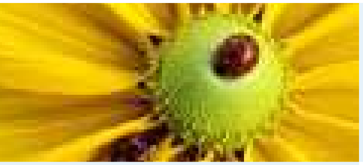
REML Estimation and Linear Mixed Models

2. Variance components models

Sue Welham

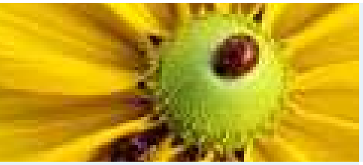
Rothamsted Research
Harpenden UK AL5 2JQ

November 18, 2008



Outline

- Issues in variance modelling
- Model determination
- Fixed or random?
- Analysis of data & use of results for designing future experiments



Variance modelling

Two schools of thought on variance modelling

- “ANOVA school”: variance model is determined by randomization procedure utilized by design
  - blocking structure from ANOVA becomes set of random terms
  - OK for designed experiments, but observational studies?
  - what about other unrandomizable terms such as time?
- variance modelling: we find the variance model that best describes the patterns of covariance within the data
  - can be used for designed experiments or observational data
  - does not respect strata in designed experiments

Why are strata important?

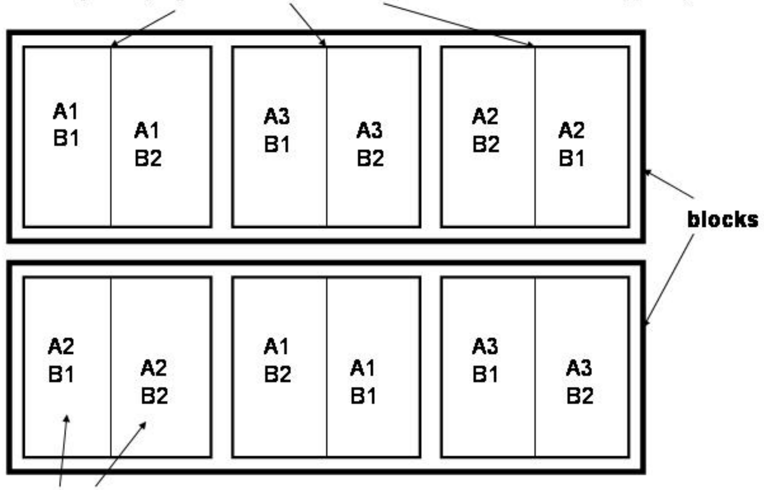
Example: split-plot design

A split-plot design is used where different experimental treatments have to be applied at different levels of experimental structure, eg

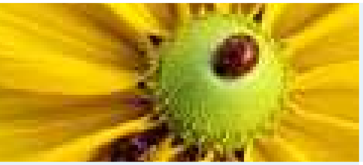
- for a field experiment looking at varietal response to irrigation: varieties can be applied to small plots, irrigation can more easily be applied to larger areas
- in CE rooms or cabinets, treatments such as temperature or elevated CO_2 levels can only be applied to whole cabinets, other treatments (variety, nutrition) may apply to individual plants

Figure 4.2: A split-plot design (block structure = blocks / whole-plots / sub-plots)

main plots (larger units to which levels of treatment A are applied)



sub-plots (smaller units to which levels of treatment B are applied)



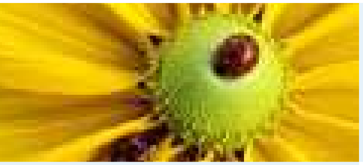
Split-plot example (2)

The split-plot design uses a nested blocking structure, with three strata (levels of experimental units)

- The largest units (blocks) are replicates of the basic design
- Within blocks, there will be several main plots (units to which treatment A is applied)
- Each main plot is split into several sub-plots (units to which the treatment B is applied

The split-plot design may be thought of as consisting of two nested RCBDs. In the first, involving variation between the large units (main plots), and the second is concerned with the variation between sub-plots within each main plot.

- The experimental units for treatment A are main plots, so variation between levels of treatment A must be compared to background variation between main plots
- The experimental units for treatment B (and the A.B interaction) are the sub-plots, and so these treatment differences must be compared to background variation between sub-plots



Split-plot example (3)

Consider a split-plot design with r blocks, w whole-plots per block and s sub-plots per whole-plot. Treatment A (w levels) is applied at random to whole-plots (within blocks) and treatment B is applied at random to sub-plots (within whole-plots).

The model for the data is written as:

$$y_{ijk} = \mu + b_i + \alpha_{s(ij)} + w_{ij} + \beta_{t(ijk)} + (\alpha\beta)_{u(ijk)} + e_{ijk}$$

where

- y_{ijk} is the observation from sub-plot k in whole-plot j in block i
- b_i is the effect of the ith block, w_{ij} is the effect of whole-plot j in block i, e_{ijk} is residual error
- with all random effects independent and $b_i \sim N(0, \sigma_b^2)$, $w_{ij} \sim N(0, \sigma_w^2)$ and $e_{ijk} \sim \sigma^2$
- μ is the grand mean
- $\alpha_{s(ij)}$ $(\sum_i \alpha_i = 0)$is the effect of the level of treatment A applied to whole-plot ij, $\beta_{t(ijk)}$ $(\sum_j \beta_j = 0)$is the effect of the level of treatment B applied to plot ijk , $(\alpha\beta)_{u(ijk)}$ $(\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0)$ is the additional effect (interaction) of the combined effect present on plot ijk

Split-plot example (4)

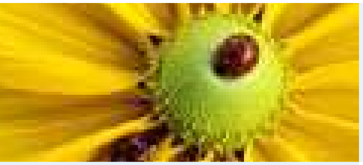
Outline
Model determination
Variance modelling
Split-plot example
Model determination
Fixed vs random
Modelling strategy
Seed weight example
References
Exercise

This is an orthogonal design and so we can construct an ANOVA table, which takes the form:

Source of variation	DF	E(MS)
Block stratum (b-1 df)		
Residual	(b-1)	$s\sigma_w^2 + \sigma^2$
Block.Wplot stratum (b(w-1) df)		
A	w-1	$f_1(\boldsymbol{\alpha}) + s\sigma_w^2 + \sigma^2$
Residual	(b-1)(w-1)	$s\sigma_w^2 + \sigma^2$
Block.Wplot.Subplot stratum (bw(s-1) df)		
B	(s-1)	$f_2(\boldsymbol{\beta}) + \sigma^2$
A.B	(w-1)(s-1)	$f_3(\boldsymbol{\alpha\beta}) + \sigma^2$
Residual	(b-1)w(s-1)	σ^2

From the point of view of randomization theory, treatment A has been randomized to whole-plots and so should be compared to background variation at the whole-plot level.

So $TMS(A)/RMS(B.W) \sim F_{(w-1),(b-1)(w-1)}$ under null hypothesis $\alpha_i = 0$ for $i = 1 \ldots w$.



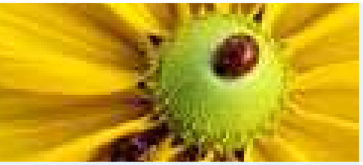
Split-plot example (5)

Now suppose $\sigma_w^2 \sim 0$ (as tested by RLRT):

- from the point of view of randomization theory, nothing changes
- a variance modelling point of view, σ_w^2 should be dropped to get a parsimonious description of the model

Does setting $\sigma_w^2 = 0$ make a difference?

- No - if the analysis is done by ANOVA, as the structure is maintained
- Yes - if a general algorithm is used that deduces structure from the variance matrix of the data
  - whole-plots are then invisible to the analysis
  - variance ratio for treatment A is compared to an F-distribution on $(w-1), (b-1)w(s-1)$ degrees of freedom, as if it had been randomized to sub-plots
  - may lead to the wrong conclusions



Model determination

Flexible recipe-based approach, based around concept of tiers (Brien & Bailey, 2006, JRSSB, also general info at http://chris.brien.name/multitier/index.html).

- Tier = set of factors with same status in the randomization
- For (one-phase) designed experiments, there are two tiers

    1. Tier 1: set of unrandomized factors (blocks)
    2. Tier 2: set of randomized factors (treatments)

    Note: randomization is regarded as giving an allocation of treatments (randomized factors) to blocks (unrandomized factors)

- Randomization links the two tiers
- Split-plot:

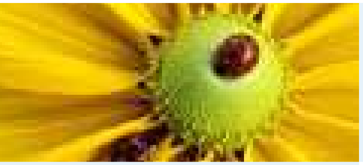
Tier 2		Tier 1
		Block
A	→	Wholeplots in Blocks
B	→	Subplots in Wholeplots in Blocks

- Randomization determines skeleton ANOVA table (sources & DF)

Model determination

Determining model recipe: mainly following Brien & Bailey (2006, JRSSB, Fig 24)

- Stage 1: randomization model
  - Determine tiers & indexing factors, eg. { Block, Wplot, Splot }, {A, B }, and links due to randomization
  - Determine intra-tier formula, eg. Block/Wplot/Splot and A*B
  - (can be helpful at this point to form skeleton ANOVA table)
- Stage 2: mixed model
  - Add inter-tier interactions to get full formulae for analysis
  - Expand formulae to get list of model terms
  - Designate model terms as fixed or random
- Stage 3: randomization-based mixed model
  - Augment the model for other terms considered important
  - Identify totally confounded terms: leave one of each set in model
  - Vary parameterization of terms, eg. modify covariance matrices
  - (may want to simplify model)



Fixed vs random

Should terms be allocated as fixed or random? Several schools of thought exist:

- School 1: randomization-based
  - Treatment terms are fixed
  - Block (and block.treatment) terms are random
- School 2: populations
  - Fixed terms represent terms with specific levels chosen for the experiment
  - Random terms represent terms where levels are a representative sample of (normal) population
- School 3: populations & prediction
  - Fixed terms as for (2) where the aim is to obtain an unbiased estimate of effects
  - Random terms as for (2), or where the aim is to obtain predictions of future performance with minimum MSE property, eg. variety selection

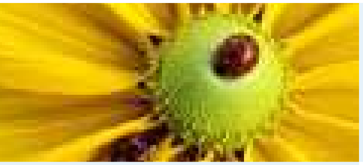
Fixed vs random

Might consider an amalgamated approach:

- School 4: randomization, populations, prediction and pragmatism
    - Terms with specific levels chosen for the experiment and the aim is unbiased estimation of effects are allocated as fixed
    - Terms with specific levels chosen for the experiment and the aim is selection are allocated as random
    - Terms associated with the randomization structure of the design (if any) are allocated as random
    - Terms whose levels are a representative sample of (normal) population (and variation is of interest) are allocated as random

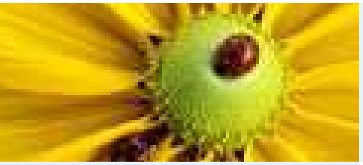
Random terms may also be used to account for correlation in the data, or to partition variability linked to treatments.

Note that interactions between fixed and random terms are automatically classified as random.



Modelling strategy

- Once we have determined the full model, we may want to simplify the model by removing unnecessary terms, using RLRT for variance model, Wald tests (or approximate F-tests) for fixed terms.
- Need to be aware of duality between fixed and random terms: variance model will try to account for systematic trend not accounted for by fixed model
- Need also to be aware of interplay between different random terms: if an important random term is omitted, the others will try to account for its variance, which may distort other parameters
- ⇒ starting from null models is not a good strategy
- Reasonable strategy:
  - Initial model contains all fixed terms, try to establish reasonable variance model (respecting randomization) and starting from full model
  - Once variance model is established, try to reduce fixed model
- Note: this corresponds to Brien & Bailey approach



Seed weight example

- Plants from lines in DH population (*B. napus*) grown for QTL study
  - DH lines: set of offspring from two very different homozygous parents which are then themselves forced to be homozygous - some genetic variation within a known background
  - try to identify points in the genome related to variation in quantitive traits of interest
- Preliminary stage: trying to identify traits showing significant variation between lines
- Glasshouse experiment with 2-3 plants of each line
- No randomization!!!! (systematic design)
- Different levels of plant structure to consider, ie. whole plant, raceme (branch), silique (pod) and individual seed
- Various traits measured
- Incomplete measurements at some levels for some variables
- We will consider analysis of the average seed weight

Data set

- Evaluation of average seed-weight per raceme
- Measured for all 95 lines on racemes 1, 3 & 4, usually on 2 plants per line, occasionally 1 plant
- Aim to quantify variation between lines
- Suspect that there might be systematic differences between seed sizes according to raceme order (raceme 1 flowers & sets seed first)

Follow recipe ⇒ determine tiers (assuming completely randomized!!)

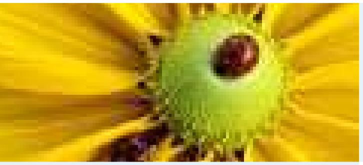
Tier 2		Tier 1
Line	→	Rep.Pot
Raceme order	--→	Raceme w/i Rep.Pot

--→ indicates not proper randomization, but the structure is correct.

Tier formulae (using standard symbolic notation) are:

- Line * Raceme = Line + Raceme + Line.Raceme
- Rep.Pot / Raceme = Rep.Pot + Rep.Pot.Raceme

Note: Rep is not a stratum because it was not used in the randomization.



Model

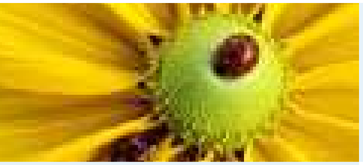
Skeleton ANOVA table (assuming 2 plants / line) with treatments as fixed terms and blocking structure as random terms:

- fixed terms = Line*Raceme
- random terms = Rep.Pot/Raceme

can be written

Source of variation	DF	EMS
Rep.Pot stratum	189	
Line	94	$f_1(\boldsymbol{\tau}) + 3\sigma_P^2 + \sigma^2$
Residual	95	$3\sigma_P^2 + \sigma^2$
Rep.Pot.Raceme stratum	380	
Raceme	2	$f_2(\boldsymbol{\tau}) + \sigma^2$
Line.Raceme	188	$f_3(\boldsymbol{\tau}) + \sigma^2$
Residual	190	σ^2

where σ_P^2 is the variance for Rep.Pot effects and σ^2 is the residual variance.



Model (2)

Outline
Model determination
Seed weight example
Data set
Model
Variance model
Computational issues
Residuals
Interpret results
Line effects
Using results
Sampling scenarios
References
Exercise

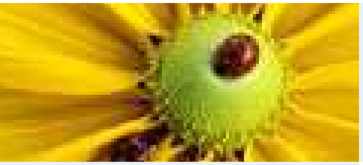
Suppose we wish to select lines ⇒ use line as random, then we have

- fixed terms = Raceme
- random terms = Rep.Pot + Line + Line.Raceme + Rep.Pot.Raceme

The skeleton ANOVA table (assuming 2 plants / line) then can be written

Source of variation	DF	EMS
Line stratum	94	
Residual	94	$6\sigma^2_L + 3\sigma^2_P + 2\sigma^2_{LR} + \sigma^2$
Rep.Pot stratum	95	
Residual	95	$3\sigma^2_P + \sigma^2$
Line.Raceme stratum	190	
Raceme	2	$f_2(\boldsymbol{\tau}) + 2\sigma^2_{LR} + \sigma^2$
Residual	188	$2\sigma^2_{LR} + \sigma^2$
Rep.Pot.Raceme stratum	190	
Residual	190	σ^2

where σ^2_L, σ^2_P, and σ^2_{LR} are the variances for Line, Rep.Pot and Line.Raceme effects, respectively and σ^2 is the residual variance.



Model (3)

Outline

Model determination

Seed weight example

Data set

Model

Variance model

Computational issues

Residuals

Interpret results

Line effects

Using results

Sampling scenarios

References

Exercise

With Line and Line.Raceme fixed, the Raceme main effect is tested against residual error:

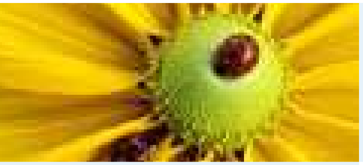
- are overall differences between Racemes large compared to within-plant variation?

With Line and Line.Raceme random, the Raceme main effect is tested against the Line.Raceme level residual variation.

- are overall differences between Racemes large compared to Line.Raceme variation? ie. is the main effect large compared to the interaction?

Which is ’right’ depends on context - the two tests answer different questions.

Here we keep Line and Line.Raceme as random as we are interested specifically in quantifying sources of variation.

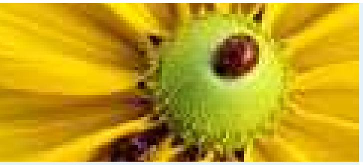


Rewriting the model

There is often more than one way of writing the model in equivalent forms. Consider two sets of factors:

Rep	Pot	Line	Plant
1	1	2	1
1	2	3	1
1	3	1	1
1	4	4	1
2	1	1	2
2	2	2	2
2	3	4	2
2	4	3	2

- Rep, Pot, Line as before, Plant labels replicate plants within lines
- Rep.Pot $\equiv$ Line.Plant
- using Line.Plant in formulae can make the model look simpler
- *but* it hides the underlying structure & randomization
- best to keep original factor set
- second best: at least start with full set of factors to get structure correct (often not clear otherwise) then revert to reduced set



Variance model

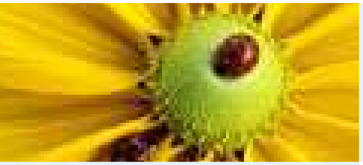
Random = Line + Line.Raceme + Rep.Pot + Rep.Pot.Raceme
Random = Line + Line.Raceme + Line.Plant + Line.Plant.Raceme

Model:

$$y_{ijk} = \mu + l_i + R_k + (lr)_{ik} + (lp)_{ij} + e_{ijk}$$

where

- y_{ijk} is the observed value for line i, plant j, raceme k
- μ is the overall mean
- l_i is the random effect of line i, with var$(l_i) = \sigma^2_L = \sigma^2\gamma_L$
- R_k is the fixed effect of raceme k
- $(lr)_{ik}$ is the effect of raceme k on line i, with var$[(lr)_{ik}] = \sigma^2_{LR} = \sigma^2\gamma_{LR}$
- $(lp)_{ik}$ is the effect of plant j of line i, with var$[(lp)_{ij}] = \sigma^2_P = \sigma^2\gamma_P$
- e_{ijk} is the residual error (Line.Plant.Raceme effects)



Variance model (2)

Covariance model $= \sigma^2 (\sum_{i=L,LR,P} \gamma_i \boldsymbol{Z}_i \boldsymbol{Z}_i' \ + \ \boldsymbol{I} \ )$

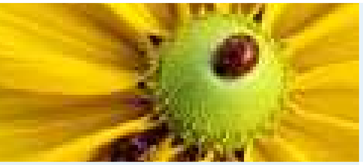
$$\operatorname{cov}\left(y_{ijk}, y_{rst}\right) = \begin{cases} \sigma^2(\gamma_L + \gamma_{LR} + \gamma_P + 1) & i = r, j = s, t = k \\ \sigma^2(\gamma_L + \gamma_P) & i = r, j = s, t \neq k \\ \sigma^2(\gamma_L + \gamma_{LR}) & i = r, j \neq s, t = k \\ \sigma^2 \gamma_L & i = r, j \neq s, t \neq k \\ 0 & i \neq r \end{cases}$$

- Overall variance matrix must remain positive-definite, eg. $\gamma_L + \gamma_{LR} + \gamma_P + 1 > 0$

- $\gamma_L \gg 0 \Rightarrow$ plants of same line more similar than plants of different line (real line differences)

- $\gamma_L \sim 0 \Rightarrow$ no real differences between lines

- $\gamma_L \ll 0 \Rightarrow$ plants of different lines more similar than plants of same line (establishment of separate lines has failed badly)



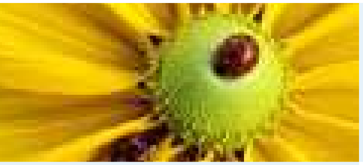
Variance model (3)

Covariance model $= \sigma^2\left(\sum_i \gamma_i \boldsymbol{Z}_i \boldsymbol{Z}_i' \; + \; 1 \right)$

$$\operatorname{cov}\left(y_{ijk}, y_{rst}\right) = \begin{cases} \sigma^2(\gamma_L + \gamma_{LR} + \gamma_P + 1) & i = r, j = s, t = k \\ \sigma^2(\gamma_L + \gamma_P) & i = r, j = s, t \neq k \\ \sigma^2(\gamma_L + \gamma_{LR}) & i = r, j \neq s, t = k \\ \sigma^2 \gamma_L & i = r, j \neq s, t \neq k \\ 0 & i \neq r \end{cases}$$

- $\gamma_P \ll 0 \Rightarrow$ seed weight less similar across racemes within plant than between plants, might indicate competition within plant
- need to allow $\gamma_L, \gamma_P < 0$
- difficult to interpret $\gamma_{LR} \ll 0$, so might constrain $\gamma_{LR} \geq 0$
- σ^2 must be positive

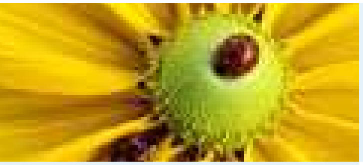
Recall: these decisions have implications for formal tests of hypotheses $\gamma_i = 0$



Computational issues

Negative variance components can cause problems with software:

- Some algorithms never form the full $(n \times n)$ variance matrix $\boldsymbol{H}$ for reasons of computational efficiency
- Checks whether $\boldsymbol{H}$ positive-definite can then only be made indirectly
- In general, no explicit form of constraints on variance components
- Can (usually) identify that $\boldsymbol{H}$ has become invalid, but not necessarily why
- If parameters go out of bounds, algorithms can fail
- Fisher scoring tends to be more stable than other gradient methods (personal experience)
- One strategy: start with components constrained positive, then re-start from that solution & allow negative components



Estimated parameters

GenStat model specification and output:

```
vcomp [fixed=Raceme] random=Line+Line.Raceme+Line.Plant/Raceme; \
                     constraint=none,pos,none,pos
reml SW
```

```
REML variance components analysis
=================================

Response variate:  SW
Fixed model:       Constant + Raceme
Random model:      Line + Line.Raceme + Line.Plant + Line.Raceme.Plant
Number of units:   567

Line.Raceme.Plant used as residual term
Sparse algorithm with AI optimisation

Estimated variance components
-----------------------------

Random term                component        s.e.
Line                        0.006905    0.001526
Line.Raceme                 0.000600    0.000269
Line.Plant                  0.004576    0.000806


Residual variance model
-----------------------

Term              Factor        Model(order)  Parameter        Estimate       s.e.
Line.Raceme.Plant               Identity      Sigma2            0.00318   0.000321
```

Estimated gammas: $\gamma_L = 2.17$, $\gamma_{LR} = 0.19$, $\gamma_P = 1.44$

Residual plots

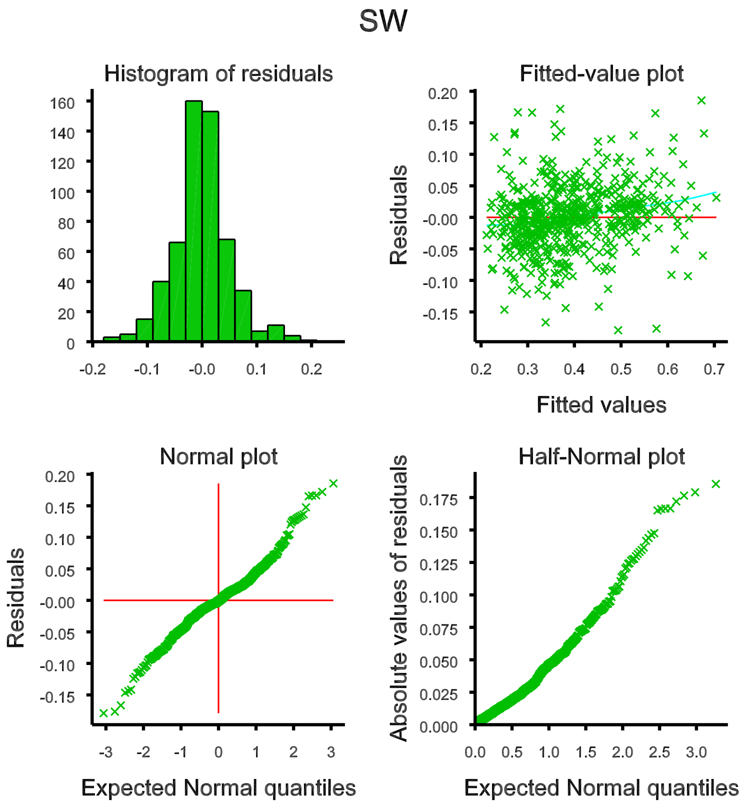
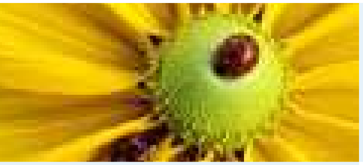


Figure 1: Residual plots for analysis of average seed weight per raceme for 95 lines

Trend in plot of fitted values vs residuals?



Residuals

Residuals are usually considered as

$$\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} - \boldsymbol{Z}\tilde{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{y}$$

with fitted values

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \tilde{\boldsymbol{e}} = \boldsymbol{X}\hat{\boldsymbol{\tau}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$$

Note: form slightly more complex in more general model - this form specific to var$(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

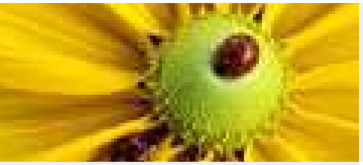
For a model defined to allow negative variance components, residuals should be considered as:

$$\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} = \boldsymbol{H}\boldsymbol{P}\boldsymbol{y}$$

as individual random terms are no longer defined, with the fitted values then consisting of just the fixed effects

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\tau}}$$

Both forms may be useful in general, depending on context.



Residuals

In the former case, consider the covariance between the fitted values and residuals:

$$\text{cov}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{e}}) = \text{cov}((\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}, \boldsymbol{P}\boldsymbol{y}) = \sigma^2(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\boldsymbol{P} = -\sigma^2(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1} \neq 0$$

Since

$$\text{cov}(\hat{\boldsymbol{\tau}}, \tilde{\boldsymbol{e}}) = 0\ ; \qquad \text{cov}(\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{e}}) = \sigma^2\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}$$

the covariance between the fitted values ($\tilde{\boldsymbol{y}}$) and residuals ($\tilde{\boldsymbol{e}}$) arises from covariance between the predictors of random effects and the residual.

This is a direct result of shrinkage and is easy to see from a simple example of a one-way random effects model:

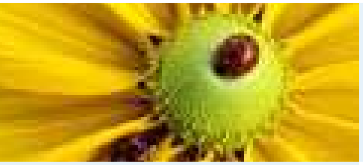
$$y_{ij} = u_i + e_{ij}$$

for $i = 1...g$, $j = 1...r$ with var$(u_i) = \sigma_u^2$ and var$(e_{ij}) = \sigma^2$, with all effects independent.

This model gives

$$\tilde{u}_i = \frac{r\gamma}{r\gamma + 1}\bar{y}_{i\cdot}$$

which is shrunk towards zero compared with the estimate $\hat{u}_i = \bar{y}_{i\cdot}$ achieved if u_i was treated as a fixed effect.



Residuals (2)

Outline
Model determination
Seed weight example
Data set
Model
Variance model
Computational issues
Residuals
Interpret results
Line effects
Using results
Sampling scenarios
References
Exercise

$$\tilde{u}_i = \frac{r\gamma}{r\gamma + 1}\bar{y}_{i\cdot}$$

The shrinkage is small when $r\gamma$ is large, ie. as replication increases or as variation in the random term increases wit respect to the residual variance. The excess is picked up by the residual term.

In this model the fitted values are the predictors of the random effects, so

$$\text{cov}(\tilde{y}_{ij}, \tilde{e}_{ij}) = \text{cov}(\tilde{u}_i, y_{ij} - \tilde{u}_i) = \sigma^2 \frac{\gamma}{r\gamma + 1} = \sigma^2 \frac{1}{r + \gamma^{-1}}$$

then using

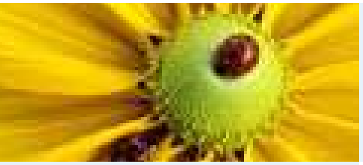
$$\text{var}(\tilde{y}_{ij}) = \sigma^2 \frac{r\gamma^2}{r\gamma + 1}, \quad \text{var}(\tilde{e}_{ij}) = \sigma^2 \frac{\gamma(r-1)+1}{r\gamma + 1}$$

we get

$$\text{corr}(\tilde{y}_{ij}, \tilde{e}_{ij}) = \frac{1}{\sqrt{r[(r-1)\gamma + 1]}}$$

The correlation in a plot of fitted values vs residuals therefore decreases as r (replication) increases and as γ increases.

This leaves a question as to whether these plots are useful.



Investigate variance model

- Of interest to test null hypothesis $\gamma_{LR} = 0$ - is there any variation in the relative raceme effects between lines (aspect of partitioning within plant)
- Increase in $-2\ell_2$ on dropping term Line.Raceme = 6.02
- Constrained term, so use 50:50 $\chi_0^2 : \chi_1^2$ mixture distribution

For random variable X with this mixture distribution and $x > 0$

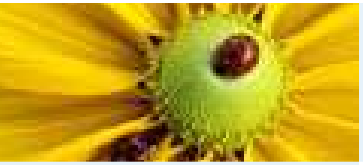
$$\begin{aligned} P(X \leq x) &= 0.5P(\chi_0^2 \leq x) + 0.5P(\chi_1^2 \leq x) \\ &= 0.5 + 0.5P(\chi_1^2 \leq x) \end{aligned}$$

so we can transform critical values of χ_1^2 distribution directly into those of mixture distribution.

Critical values:

	P($X \leq x$)			
x	0.90	0.95	0.99	0.995
50:50 mixture	1.64	2.67	5.32	6.80

But this is an asymptotic test, so it may (in general) be useful to check the empirical distribution of the tests statistic under the null hypothesis using a parametric bootstrap.



Parametric bootstrap

The model for this data takes the form:

$$y_{ijk} = \mu + l_i + R_k + (lr)_{ik} + (lp)_{ij} + e_{ijk}$$

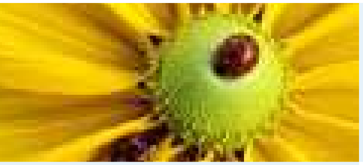
To evaluate the distribution of the RLRT statistic for null hypothesis $\gamma_{LR} = 0$ means generating simulated data y^*_{ijk} as

$$y^*_{ijk} = \hat{\mu} + l^*_i + \hat{R}_k + ((lp)^*_{ij} + e^*_{ijk}$$

where $l^*_i \sim N(0, \hat{\sigma}^2_L)$, $(lp)^*_{ij} \sim N(0, \hat{\sigma}^2_P)$ and $e^*_{ijk} \sim N(0, \hat{\sigma}^2)$, using estimated parameters from the full model.

Generate n data sets (n large) and perform the RLRT for null hypothesis $\gamma_{LR} = 0$ for each data set, by fitting the model including γ_{LR} (constrained positive), and omitting γ_{LR}.

The empirical distribution of the RLRT can be used to indicate critical values for the original tests statistic.



Parametric bootstrap (2)

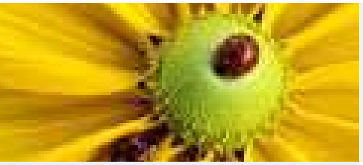
In this case, critical values are generated as:

	P($X \leq x$)			
x	0.90	0.95	0.99	0.995
50:50 mixture	1.64	2.67	5.32	6.80
Simulation (10000 runs)	1.64	2.71	5.41	6.64

Here, the closeness of the empirical distribution to the expected χ^2 mixture is unsurprising as there is no shortage of information (cf ANOVA table: $\sim$ 188 df for Line.Raceme SS).

In smaller data sets, test statistics may be less well-behaved. So we rejectthe null

hypothesis $\gamma_{LR} = 0 \Rightarrow$ Line.Raceme interaction is present (but small compared to Line and Line.Plant variation).



Investigate model

Looking at the remaining random terms:

- Dropping Line from model gives increase in $-2RL$ of 32.99, indicating real differences between lines
- We do not drop Rep.Pot (Line.Plant) from model as this is part of randomization structure and forms residual SS for Line comparisons

Having decided to keep all terms in random model, consider fixed term (Raceme)

- This has approximate F-statistic 10.96 with denominator DF = 182.9 (p<0.001)
- Note: denominator DF changes to 370.7 if Line.Raceme term dropped from model
- Predicted means (×100): Raceme 1=0.411, 3=0.382, 4=0.384, with Ave SED = 0.006
- ⇒ seeds on raceme 1 tend to be larger, with some small variation (SD=0.024) in differences across lines

Line effects

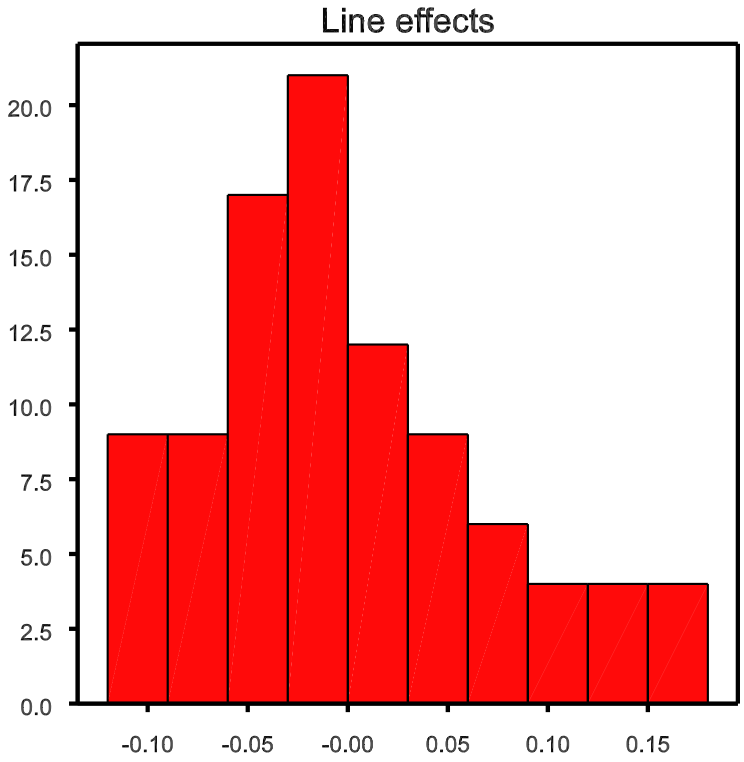
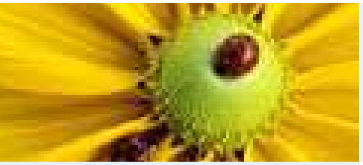


Figure 2: Histogram of line effects



Using results

We want to use these results to help design sampling strategies in future experiments on the same population.

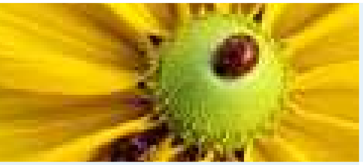
- *e.g.* what size sample is required to detect whether there are differences between lines in average seed weight at a given level of plant structure (whole plant, raceme, silique or individual seed)

- Model for variation for individual seed weight (ignoring Line.Raceme effects for simplicity):

$$y_{ijklm} = \mu + R_k + \ell_i + p_{ij} + r_{ijk} + s_{ijkl} + e_{ijklm}$$

  where

  - y_{ijklm} is weight of mth seed in lth silique on kth raceme of jth plant of line i
  - μ and R_k fixed effects, all other effects random
  - l = line, p = plant w/i line, r = raceme w/i plant, s = silique w/i raceme, e = seed w/i silique
  - $\text{var}\left(\ell_i\right) = \sigma^2_\ell$, $\text{var}\left(p_{ij}\right) = \sigma^2_P$, $\text{var}\left(r_{ijk}\right) = \sigma^2_R$, $\text{var}\left(s_{ijkl}\right) = \sigma^2_s$, $\text{var}\left(e_{ijklm}\right) = \sigma^2_e$



Using results

For raceme level data, with data measured from N_R racemes on N_P plants per line, if we consider the full structure of the plant, then average seed weight per raceme then takes the (approximate) form

$$\bar{y}_{ijk..} = \mu + R_k + \ell_i + p_{ij} + r_{ijk} + \frac{1}{N_s} \sum_{l} s_{ijkl} + \frac{1}{N_s N_e} \sum_{lm} e_{ijklm}$$

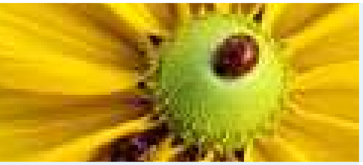
with

$$\mathrm{var}\left(\bar{y}_{ijk..}\right) = \sigma_\ell^2 + \sigma_P^2 + \sigma_R^2 + \frac{\sigma_s^2}{N_s} + \frac{\sigma_e^2}{N_s N_e}$$

where

- N_s is the number of siliques per raceme
- N_e is the number of seeds per raceme
- and it is assumed that the structure is balanced

The residual variance for a raceme level variable is thus a composite of the components σ_R^2, σ_s^2, σ_e^2 and the plant structure (N_s, N_e).



Using results

Similarly, line means take the (approximate) form

$$\begin{aligned}\bar{y}_{i\ldots.} &= \mu + \frac{1}{N_R}\sum_k R_k + \ell_i + \frac{1}{N_P}\sum_j p_{ij} + \frac{1}{N_P N_R}\sum_{jk} r_{ijk} \\ &+ \frac{1}{N_P N_R N_s}\sum_{jkl} s_{ijkl} + \frac{1}{N_P N_R N_s N_e}\sum_{jklm} e_{ijklm}\end{aligned}$$

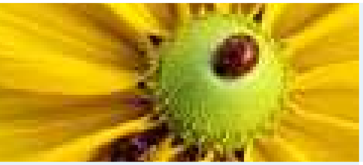
with

$$\operatorname{var}\left(\bar{y}_{i\ldots.}\right) = \sigma_\ell^2 + \frac{\sigma_P^2}{N_P} + \frac{\sigma_R^2}{N_P N_R} + \frac{\sigma_s^2}{N_P N_R N_s} + \frac{\sigma_e^2}{N_P N_R N_s N_e}$$

The variation of line effects as a proportion of the total variation of line means is

$$\sigma_\ell^2 / \operatorname{var}\left(\bar{y}_{i\ldots.}\right)$$

If our aim is to detect line differences, then we want to structure our sample to maximise the proportion of variation attributable to line variation.

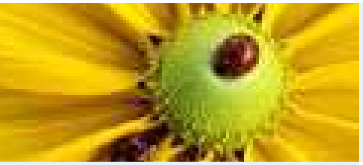


Using results

We can establish σ_s^2 and σ_e^2 from other data sets to get

σ_ℓ^2	σ_P^2	σ_R^2	σ_s^2	σ_e^2
.0069	.0046	.0031	.0012	.0050
γ_ℓ	γ_P	γ_R	γ_s	γ_e
1.38	0.92	0.62	0.24	1

- Clearly seed-to-seed variation within silique (σ_e^2) is relatively large, but we will average over lots of seeds, so not a problem
- Other large source of variation is plant-to-plant variation (σ_P^2). Usually we do not sample too many plants per line, so this is more of a problem.
- Also need to remember that we want to sample as many lines as possible so that we get both a good estimate of the line variance, and a representative sample of population



Scenario 1

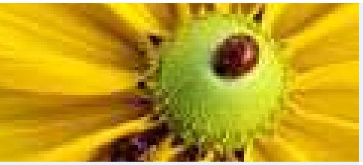
First, we consider a balanced scenario

- Sample: 30 lines, 3 plants/line, 2 racemes/plant, 2 siliques/raceme, 10 seeds/silique
- Total siliques = 360

Skeleton ANOVA table

Source	Units	DF
Line	30	29
Line.Plant	90	60
Line.Plant.Raceme	180	90
Line.Plant.Raceme.Silique	360	180

$$\begin{aligned} \text{var}\left(\bar{y}_{i....}\right) &= \sigma_\ell^2 + \frac{\sigma_P^2}{3} + \frac{\sigma_R^2}{6} + \frac{\sigma_s^2}{12} + \frac{\sigma_e^2}{120} \\ &= \sigma_e^2(1.38 + 0.44) \end{aligned}$$



Scenario 2

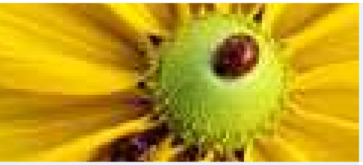
Second, consider an unbalanced scenario

- Sample 90 lines, divided into 3 sets
- Set 1: 30 lines, 3 plants, 1 raceme, 1 silique, 10 seeds
- Set 2: 30 lines, 2 plants, 1 raceme, 2 siliques, 10 seeds
- Set 3: 30 lines, 2 plants, 2 racemes, 1 silique, 10 seeds
- Total siliques = 330

Skeleton ANOVA table

Source	Units	DF
Line	90	89
Line.Plant	210	120
Line.Plant.Raceme	270	60
Line.Plant.Raceme.Silique	330	60

$$\text{var}\left(\bar{y}_{i....}\right) = \begin{cases} \sigma_e^2(1.38 + 0.63) & \text{Set 1} \\ \sigma_e^2(1.38 + 0.86) & \text{Set 2} \\ \sigma_e^2(1.38 + 0.70) & \text{Set 3} \end{cases}$$

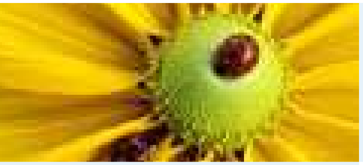


Scenario 1 vs Scenario 2

Scenario 1	Scenario 2
Balanced	Unbalanced
360 siliques	330 siliques
ANOVA analysis	REML analysis
30 lines sampled	90 lines sampled
h^2=0.86	h^2=0.62-0.69

where h^2 represents the proportion of total variation due to lines

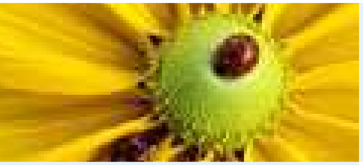
- In scenario 1, estimates of variation at all levels are estimated across all lines - but only 30 lines sampled
- In scenario 2, raceme variability estimated only in set 3 (30 lines), silique variability estimated only in set 2 (30 lines) - but 90 lines sampled in total
- If good information on variance estimates is available it may not be necessary to re-estimate variation from different levels, then random (or structured) sample of N seeds per plant can be used



References

Bailey, R.A. (2008) *Design of Comparative Experiments* Cambridge University Press, Cambridge.

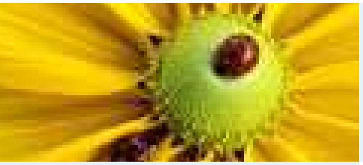
Brien CJ & Bailey RA (2006) Multiple randomizations. *Journal of the Royal Statistical Society, Series B*, **68**, 571-599.



Exercises

1. Seed variability

- Data set seedsize.xls arose from a pilot study for another Brassica species.
- Aim is to identify important sources of variation for studying average seedsize at the raceme level.
- Data sampled was 4 lines, with 4 plants per line and 3 racemes per plant.
- Establish a suitable model(s) using REML analysis and compare this to ANOVA analysis
- Points of interest:
  - adequacy of Wald test(s)
  - adequacy of mixture distribution for testing random effects constrained positive
- Program file (seedsize.gen) provided for those not familiar with GenStat



Exercises (2)

2. Sources of variation in an industrial process

- The data in the example were obtained to investigate sources and sizes of variability in an industrial process, the production of car voltage regulators (Example S from Cox and Snell 1981). Within the factory, each regulator was passed from the production line to a setting station where it was adjusted to operate within the correct range of voltages. It would then be passed to a testing station where it would be tested and sent back if outside the acceptable range. An experiment was designed to examine the sources of variability in the voltages produced by the regulators. This experiment used four testing stations, and ten setting stations: between four and eight regulators from each setting station were tested on all four testing stations.
- The data set is held in file voltage.xls.
- Fit a mixed model to determine the major sources of variation in this data.