ASReml workshop

2.1 A matrix refresher

Arthur Gilmour

NSW DEPARTMENT OF
PRIMARY INDUSTRIES

Some vectors

- A *vector* is a column of numbers
  $\boldsymbol{y}$ is the response variable
  $\hat{\boldsymbol{\tau}}$ is the fixed effects
  $\tilde{\boldsymbol{u}}$ is the random effects
  $\tilde{\boldsymbol{\epsilon}}$ is the residuals

Some matrices

- A *matrix* is a rectangular array of numbers
  $\boldsymbol{X}$ is the design matrix for fixed effects
  $\boldsymbol{Z}$ is the design matrix for random effects
  $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$ is the whole design matrix
  $\boldsymbol{G} =$var$(\boldsymbol{u})$
  $\boldsymbol{R} =$var$(\boldsymbol{\epsilon})$
  $\boldsymbol{A}$ is a relationship matrix

Matrix operations

- Transpose

$$[a\ b\ c]' = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

- Addition - matrices of the same order are added element by element

$$[1\ \ 2] + [3\ \ 4] = [4\ \ 6]$$

- Multiplication by scalar

$$3[1\ \ 2] = [3\ \ 6]$$

Matrix operations

- Matrix multiplication – must be conformable

$$\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \begin{pmatrix} A \\ B \\ C \end{pmatrix} = \begin{pmatrix} aA + bB + cC \\ dA + eB + fC \end{pmatrix}$$

- Direct product

$$\begin{pmatrix} a \\ b \end{pmatrix} \otimes [A\ B\ C] = \begin{pmatrix} aA & aB & aC \\ bA & bB & bC \end{pmatrix}$$

Mixed model equations

- Mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

REML

- Let $\boldsymbol{C} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix}$

$$\boldsymbol{P} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}$$

$$l_R = -(\log|\boldsymbol{C}| + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \nu\log\sigma^2 + \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^2)/2$$

Two forms

- $\boldsymbol{V} = \sigma^2(\boldsymbol{R} + \boldsymbol{ZGZ}')$
  default for univariate single site analyse.
  $\boldsymbol{R}$ defined as a correlation matrix
  $\boldsymbol{G}$ defined as a variance ratio
- $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ}'$
  used for multivariate and multisite analyses.

Differentiation

- $l_R = -(\log|\boldsymbol{C}| + \log|\boldsymbol{R}| + \log|\boldsymbol{G}| + \nu \log\sigma^2 + \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma^2)/2$
- $\partial l_R/\partial\phi_i = -(-\boldsymbol{C}_i\boldsymbol{C}^{Z'Z} + 0 + \mathrm{tr}(\boldsymbol{G}_i\boldsymbol{G}^{-1}) + 0 - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}_i/\sigma^2)/2$
  $\boldsymbol{y}_i = \boldsymbol{Z}\boldsymbol{G}_i\boldsymbol{G}^{-1}\tilde{u}$
- $\partial l_R/\partial\kappa_j = -(-\boldsymbol{C}_j\boldsymbol{C}^{-1} + \mathrm{tr}(\boldsymbol{R}_j\boldsymbol{R}^{-1}) + 0 + 0 - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}_j/\sigma^2)/2$
  $\boldsymbol{y}_j = \boldsymbol{R}_j\boldsymbol{R}^{-1}e\tilde{t}a$

The challenge

- to define $\boldsymbol{X}$ $\boldsymbol{Z}$ $\boldsymbol{R}$ and $\boldsymbol{G}$ to obtain the desired analysis.
- Often several ways of writing equivalent models.

ASReml workshop

2.2 Sire and Animal model

Arthur Gilmour

NSW DEPARTMENT OF
PRIMARY INDUSTRIES

Harvey.dat

- animal sire dam
  line DamAge ADG Age WT

```
101 1 0 1 3 192 390 224 1
102 1 0 1 3 154 403 265 1
103 1 0 1 4 185 432 241 1
104 1 0 1 4 183 457 225 1
105 1 0 1 5 186 483 258 1
106 1 0 1 5 177 469 267 1
```

Harvey.dat summary

- 9 sires representing 3 sire lines
  (1 2 3) (4 5) (6 7 8 9)
  8 8 5 8 7 6 8 7 8 records (65)
- Data set originally distributed with Harvey's program

2.2 Sire model

- Harvey Test data - Sire Model

```
animal sire 9 dam line 3
DamAge ADG Age WT
harvey.dat
ADG ~ mu line DamAge !r sire
```

Summary

Model term	Size	#mv	#00	MinNon0	Mean	MaxNon
1 ID		0	0	101.0	133.0	165.
2 Sire	9	0	0	1	5.0154	
3 Dam		0	65	0.000	0.000	0.00
4 Line	3	0	0	1	2.1231	
5 DamAge		0	0	3.000	4.385	5.00
6 ADG Variate		0	0	144.0	176.6	206.
7 Age		0	0	337.0	416.8	498.
8 InitialWT		0	0	144.0	241.1	300.

Fixed Design

- $\boldsymbol{X}$ has 5 columns
  `mu` is a column of ones
  `line-1` has 21 1's, 15 0's, 29 0's
  `line-2` has 21 0's, 15 1's, 29 0's
  `line-3` has 21 0's, 15 0's, 29 1's
  `DamAge` (covariate) has vector of dam ages
- This design has 1 singularity because the three `line` columns sum to give the `mu` column.
- ASReml will set $\tau_2 = 0$

Random design

- $\boldsymbol{Z}$ has 9 columns being zeros except for 8 8 5 8 7 6 8 7 8 1's respectively indicating which sire
- The variance model is $\sigma_e^2(I + \gamma \boldsymbol{Z}\boldsymbol{Z}')$ where $\sigma_s^2 = \gamma\sigma_e^2$
- The genetic model
  $\sigma_A^2 = 4\sigma_s^2;\ \sigma_s^2 = 0.25\sigma_A^2$
  $\sigma_E^2 = \sigma_e^2 - 3\sigma_s^2;\ \sigma_e^2 = \sigma_E^2 + 0.75\sigma_A^2$

Components

- 5 LogL=-188.777 S2= 132.76 61 df 0.2176 1.000

Source	terms	Gamma	Component	Comp/SE	% C
Sire	9	9	0.217651	28.8946	1.04	0 P
Variance	65	61	1.00000	132.756	5.25	0 P

- Notice $\gamma = 0.21765$, $\sigma_e^2 = 132.756$, $\sigma_s^2 = \gamma\sigma_e^2 = 28.8946$.

Genetic components

- Create a PIN file (harvey.pin)

```
#                    1 is Sire component
#                    2 is Residual
F PhenVar 1 2     #3 is Sire + Residual
F GenVar 1*4.     #4 is Sire x 4.0
H 4 3             #Heritability is GenVar/Phe
```

- Run using `ASReml -p harvey`
  This extracts the variance components from the .asr file and their variances from the .vvp file and computes the requested quantities

Genetic parameters

```
   3 PhenVar  1      161.7     35.07
   4 GenVar   1      115.6     110.8
Heritability = GenVar 4/PhenVar 3
             =      0.7150   0.5877
Notice: The parameter estimates are
        followed by their
        approximate standard errors
```

Sire BLUPS

- from the .sln file – sum to zero within lines

```
Sire    1     -0.5386E-01    4.137
Sire    2       3.591        4.142
Sire    3      -3.538        4.269
Sire    4      -5.112        4.468
Sire    5       5.112        4.468
Sire    6     0.6051E-01     4.055
Sire    7       3.420        3.914
Sire    8      0.4042        3.990
Sire    9      -3.885        3.914
```

ANOVA

- Degrees of Freedom and Stratum Variances

```
      5.93    341.034          7.2      1.0
     55.07    132.756          0.0      1.0

ANOVA          NumDF    DenDF    F-incr      Prob
  9 mu             1      5.9   5906.95     <.001
  4 Line           2      5.9      6.19     0.035
  5 DamAge         1     57.8      0.62     0.435
WARNING: The DenDF values are calculated ig
         fixed/boundary/singular variance p
         and may change with the order of f
```

- Damage is NS; Line is significant tested against Sire variance

Fixed effects

Term	Level	Effect	SE
DamAge	1	-1.478	1.881
Line	1	0.000	0.000
Line	2	-14.41	6.286
Line	3	6.466	5.293
mu	1	183.5	9.319

- Notes:
  Terms are in reverse order (ASReml solves from bottom)
  Line-1 effect is singular.

Predict

- predict Line

```
Predicted values of ADG
DamAge evaluated at average value 4.
Sire   is ignored in the prediction

Line  Predicted_Value Stand_Error Ec
 1.0000       177.0109       4.0193 E
 2.0000       162.5983       4.8301 E
 3.0000       183.4766       3.4415 E
Overall Stnd Error of Diff 5.851
```

Line+Sire

- predict Sire !present Sire Line

```
Line is average of combinations present
DamAge  is evaluated at    4.3846

 Sire    Pred_Value Stand_Error Ecode
 1.0000   176.9570      3.5634 E
 2.0000   180.6023      3.5962 E
 3.0000   173.4734      4.1885 E
 4.0000   157.4867      3.6989 E
 5.0000   167.7099      3.8875 E
 6.0000   183.5371      3.8512 E
```

Line+Sire

- predict Sire !present Sire Line

```
   3.0000    173.4734        4.1885 E
   4.0000    157.4867        3.6989 E
   5.0000    167.7099        3.8875 E
   6.0000    183.5371        3.8512 E
   7.0000    186.8966        3.4811 E
   8.0000    183.8807        3.6774 E
   9.0000    179.5918        3.4811 E
 Overall Stnd Error of Diff 5.146
```

ASReml workshop

2.3 Animal model

Arthur Gilmour

NSW DEPARTMENT OF
PRIMARY INDUSTRIES

Relationship Matrix

- The diagonal of the relationship matrix ($a_{i,i}$) is $1 + f$ where $f = a_{s,d}/2$ is the inbreeding coefficient.
- The relationship of an animal with other animals is the average of its parental values.
- When parents are listed before progeny, this gives a straight forward way to calculate relationships.
- However, the inverse relationship matrix is in fact easier to calculate and more sparse.

reducing to

- $\boldsymbol{A}_i^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{p}q\boldsymbol{p}' & \boldsymbol{p}q \\ q\boldsymbol{p}' & q \end{pmatrix}$ where $\boldsymbol{p}$ is zero except for two $\frac{1}{2}$'s in parental rows, $q = (1 + f_i - \boldsymbol{p}'\boldsymbol{A}\boldsymbol{p})^{-1} = (1 - (a_{s,s} + a_{d,d})/4)^{-1}$ so that it just requires keeping diag($\boldsymbol{A}$).
- For the case of a sire model (dams unknown, sires not inbred),

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{I}_9 & 0.5\boldsymbol{Z}_s' \\ 0.5\boldsymbol{Z}_s & 0.25\boldsymbol{Z}_s\boldsymbol{Z}_s' + 0.75\boldsymbol{I}_{65} \end{pmatrix}$$

Variance model

- Fitting an animal model for the harvey data
  $\boldsymbol{V} = \sigma^2_E \boldsymbol{I}_{65} + \sigma^2_A \boldsymbol{Z} \boldsymbol{A} \boldsymbol{Z}'$
  where $\boldsymbol{Z} = (\boldsymbol{0}_{65\times 9}\ \boldsymbol{I}_{65})$
  Thus $\boldsymbol{V}$ simplifies to
  $\sigma^2_E + \sigma^2_A(0.25\boldsymbol{Z}_s\boldsymbol{Z}'_s + 0.75\boldsymbol{I}_{65})$
  $= (\sigma^2_E + 0.75\sigma^2_A)\boldsymbol{I}_{65} + 0.25\sigma^2_A\boldsymbol{Z}_s\boldsymbol{Z}'_s$ which is exactly the same as under the sire variance model.

Running the model in ASReml

```
!REDO !ARG !? 1=Sire 2=Animal
Harvey Test data - Sire and Animal M
 animal !P sire 9 dam line 3 # Added
 DamAge ADG Age WT
harvey.dat     # Pedigree file line a
harvey.dat     !DOPART $1 # Data
!PART 1
 ADG ~ mu line DamAge !r sire
!PART 2
 ADG ~ mu line DamAge !r animal
```

Results - components

```
5 LogL=-188.781 S2=54.092 61df 1.957
6 LogL=-188.777 S2=47.314 61df 2.411
7 LogL=-188.777 S2=46.100 61df 2.506
Final parameter values          2.508

Source    terms  Gamma Component C/SE
animal     74 74 2.5086  115.578 1.04
Variance   65 61 1.0000  46.0723 0.51
```

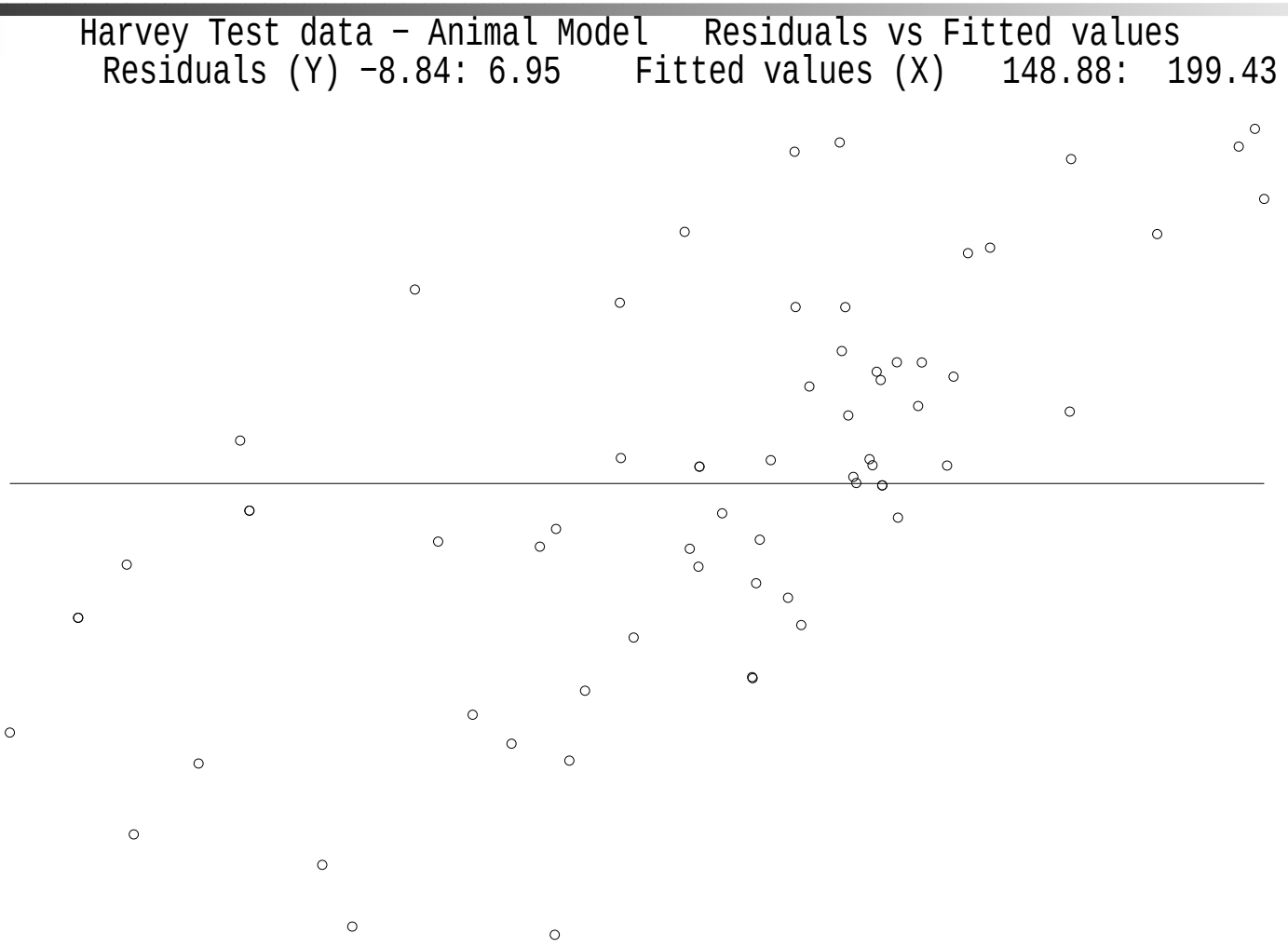
Results - ANOVA

- ANOVA NumDF DenDF F-incr Prob
  9 mu 1 5.9 5906.95 <.001
  4 line 2 5.9 6.19 0.035
  5 DamAge 1 57.8 0.62 0.435
  is the same as the Sire model

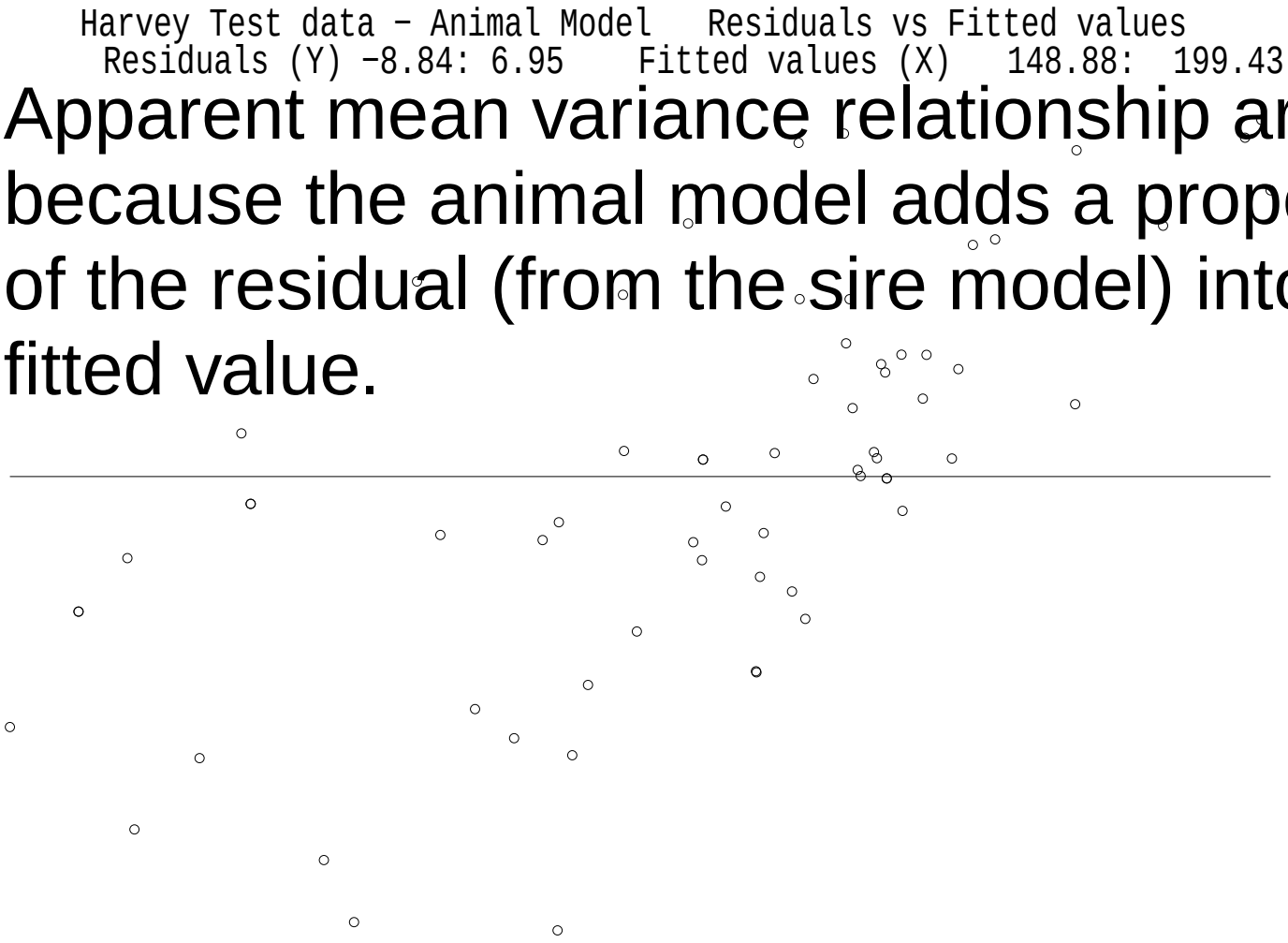
Results - Effects

		Estimate	StandErr	T-val
5 DamAge	1	-1.47752	1.88080	-0.79
4 line	2	-14.4126	6.28603	-2.29
	3	6.46567	5.29341	1.22
9 mu	1	183.489	9.31871	19.69
1 animal		74 effects fitted		

Plot of Residuals



Plot of Residuals



- Apparent mean variance relationship arises because the animal model adds a proportion of the residual (from the sire model) into the fitted value.

Other genetic components

- Genetic maternal definition `DAM !P` and include DAM in the model
- Permanent environment effect Use `ide(animal)`
- Maternal environment effect Use `ide(DAM)`

Estimability of Components

- Hendersons method III equated the Sire Mean Square (SMS) to its expectation $\sigma_e^2 + k\sigma_s^2$
- While SMS ≥ 0, $\sigma_s^2 =$ (SMS-EMS)$/k$ is $\geq$ -EMS$/k$
- In REML, often constrain components to be positive (equivalent to estimating σ_e^2 after dropping `sire` from the model).

Sire model $\rightarrow$ Animal model

- Solving the model in terms of genetic components can lead to further problems. $\sigma^2_E = \sigma^2_e - 3\sigma^2_s$ may be negative and not estimable (ASReml requires a positive residual variance)
- $h^2 = 4\sigma^2_s/(\sigma^2_e + \sigma^2_s)$ is ≤ 4 but if the genetic model is correct, should be less than 1.
- I.e. Our model may not adequately represent the variation in the data leading to unacceptable genetic parameter estimates.

ASReml workshop

2.4 Bivariate Analysis

Arthur Gilmour

NSW DEPARTMENT OF
PRIMARY INDUSTRIES

Bivariate Analysis

- With Harvey data, could analyse ADG and WT - to get two sets of sire effects and two sets of residuals.
- Consider sire effects ($\boldsymbol{U}$) in a 9×2 table. Rows are independent: $\Sigma_R = \boldsymbol{I}_9$ Columns are correlated but have different variances: $\Sigma_C = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$

Bivariate continued

- Now stack WT effects below ADG effects:
  $\boldsymbol{u} = \text{vec}(\boldsymbol{U})$
  $\text{var}(\boldsymbol{u}) = \Sigma_C \otimes \Sigma_R$
- Essential to get order correct: second term is nested within first.
- Adding part 3 to the Harvey job: the model

```
!PART 3
ADG WT ~ Trait Tr.line,
         at(Tr,2).Age !r Tr.sire
```

Variance header line

```
ADG WT ~ Trait Tr.line,
         at(Tr,2).Age !r Tr.sire
1 2 1                  # <============
0    # 65 records
Trait 0 US
3*0
Tr.sire 2
Tr 0 US
3*0
sire 0 ID
```

1 R structure

- ```
  ADG WT ~ Trait Tr.line,
           at(Tr,2).Age !r Tr.sire
  1 2 1
  0    # 65 records # <============
  Trait 0 US         # <============
  3*0                # <============
  Tr.sire 2
  Tr 0 US
  3*0
  sire 0 ID
  ```
  ```

1 G Structure

- ADG WT ~ Trait Tr.line,
  at(Tr,2).Age !r Tr.sire

```
1 2 1
0    # 65 records
Trait 0 US
3*0
Tr.sire 2          # <============
Tr 0 US            # <============
3*0                # <============
sire 0 ID          # <============
```

Notes

- R structure is 65×2 because data is ordered traits within records
- G structure is 2×9 because `Tr.sire` means the effects are ordered sires within traits
- Initial values must be supplied but are often difficult to guess. Specifying them as zeros (`3*0`) tells ASReml to work out some values from the data.

Missing data

- Multivariate analysis, with US error variance matrix, ASReml will automatically handle missing values in the traits.

Output:convergence

- ```
  1 LogL=-466.835 S2=1.0000 123 df
  2 LogL=-451.914 S2=1.0000 123 df
  3 LogL=-437.150 S2=1.0000 123 df
  4 LogL=-428.395 S2=1.0000 123 df
  5 LogL=-427.247 S2=1.0000 123 df
  6 LogL=-427.201 S2=1.0000 123 df
  7 LogL=-427.201 S2=1.0000 123 df
  ```
- Analysis on Variance scale
  $\boldsymbol{V} = \sigma^2(\boldsymbol{R} + \boldsymbol{ZGZ}')$ with $\sigma^2 = 1$.
  ```

Output:components

Source	Model	terms	Gamma	Component	C/SE	%	C
Residual	UnStru	1 1	132.370	132.370	5.29	0	U
Residual	UnStru	2 1	-98.0288	-98.0288	-2.27	0	U
Residual	UnStru	2 2	659.005	659.005	5.19	0	U
Tr.sire	UnStru	1 1	27.2002	27.2002	1.02	0	U
Tr.sire	UnStru	2 1	-12.1986	-12.1986	-0.30	0	U
Tr.sire	UnStru	2 2	98.7165	98.7165	0.88	0	U

Output:structures

- ```
  Covar/Var/Corr UnStructured
   132.4     -0.3319
  -98.03       659.0
  Covar/Var/Corr  UnStructured
   27.20     -0.2354
  -12.20       98.72
  ```
  ```

Output:ANOVA

- ANOVA NumDF DenDF F-incr Pro
  9 Trait 2 4.9 5019.79 <.00
  10 Tr.line 4 5.3 2.65 0.15
  12 at(Tr,2).Age 1 58.5 6.33 0.01

Usual process

- Univariate analysis of each trait.
  - Identify an appropriate fixed model
  - Check for outliers and problems with data structure
  - Sort out fixed model appropriate for each trait
  - Ensure there is variance at each level: if there is no sire variance in a univariate model, ASReml will not be able to estimate it in a multivariate model

Multivariate analysis

- Modelling becomes more difficult as number of traits increases - there may be variance but the matrix may not be positive definite (covariances too big).
  - Maybe do bivariate pairs first
  - !GP constrain positive definite - will use EM updates if AI updates generate NPD matrix
  - try FA or CHOL reduced parameterization
  - try Singular XFA or CHOL paramaterization

Exercises

- Coopworth data set - see Reference manual
- Five traits with varying amounts of data.
- No depth of pedigree (dams not linked to sires)
- Do univariate analyses
- Do bivariate analyses.
- Use COOP data set and attempt multivariate models.