NC STATE
UNIVERSITY

Linear Mixed Models - A Short Review

C. Maltecca

Quantitative Genomics and Modern Breeding: from theory to practice.
Florence Aug 26-30, 2024

Mixed Models in a Nutshell: Theory and Concepts

Mixed linear models are a particular class of models containing both fixed and random effects. Loosely speaking, a mixed model is a model where some terms remain constant over repeated sampling and some other terms vary at random according to some distribution. For simplicity from now on we will drop the notation linear and refer to these models simply as mixed models. For any mixed model we can identify three main components: the equation of the model, the expectations and Variance-Covariance for the random effects, and all the remaining assumptions regarding the model.

The model

A model is a mathematical representation of our understanding of the biological process that explains our observations. We can think of each observation as a single equation (and in this case we are confining ourselves to linear equations), containing the trait of interest, and the factors that explain the observations. For example:

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$$

A system of equations is then a set of these single linear equations and a solution for the system must satisfy all equations. With *n* unknown the system takes form:

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\
a_{31}x_1 + a_{32}x_2 + \ldots + a_{3n}x_n &= b_3 \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m
\end{aligned}$$

The model

The same set of equation can be rewritten in a more convenient matrix notation

$$\mathbf{Ax} = \mathbf{b}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} + a_{12} + \ldots + a_{1n} \\ a_{21} + a_{22} + \ldots + a_{2n} \\ a_{31} + a_{32} + \ldots + a_{3n} \\ \vdots \\ a_{m1} + a_{m2} + \ldots + a_{mn} \end{bmatrix}, \ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

The model

From now on we will assume that our model contain both fixed and random effects (more on it below). Traditionally mixed models have been represented in matrix form as follow

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where
y is the vector of the observations (observables),
b is a vector of fixed effects
u is a vector of random effects (unknown),
e is a vector of residual (whatever we cannot explain with our model) which are also random,
X and **Z** are incidence matrices, assigning each element of **b** and **u** to their corresponding element in **y**.

Fixed and random effects

The distinction between fixed and random applies to the unknown model components.

A fixed effect is a known constant that will remain the same over conceptual repeated sampling

A random effect is a random variable that arises from the subsampling and random selection of "treatment" levels.

In reality the distinction between fixed and random effects often depends on the practical use and interpretation of parameter estimates. When the investigator is interested in comparing specific levels of a certain factors (let's say amount of fertilizer for a plant or concentrate for a cow) than it is sensible to consider them as fixed effect. When a parameter is not of relevance for the analysis but rather a nuisance that we want to account for, more often than not we end up treating that effect as random. Also keep in mind that one of the advantages of random effects is their parsimony.

Expectations and variance-covariance for the random effects

Since we have assumed that random effects come from some large population we need to define location and dispersion for these parameters. Note that we have assumed that our observations are normally distributed and we will always assume that our residuals are normally distributed. Keeping the general matrix notation that we have seen before the expectations of, **u** **e** and **y** are

$$E(\mathbf{u}) = \mathbf{0}$$

$$E(\mathbf{e}) = \mathbf{0}$$

$$\begin{aligned} E(\mathbf{y}) &= E(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \\ &= E(\mathbf{Xb}) + E(\mathbf{Zu}) + E(\mathbf{e}) \\ &= \mathbf{X}E(\mathbf{b}) + \mathbf{Z}E(\mathbf{u}) + E(\mathbf{e})\, , \\ &= \mathbf{Xb} + \mathbf{0} + \mathbf{0} \\ &= \mathbf{Xb} \end{aligned}$$

Expectations and variance-covariance for the random effects

Also, the variances of **u** and **e** are

$$\mathbf{V} = \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$$

The actual structure of **G** and **R** is flexible we will discuss this in later lectures but in the simplest cases

$$\mathbf{G} = \mathbf{I}\sigma_u^2 \text{ and } \mathbf{R} = \mathbf{I}\sigma_e^2$$

Expectations and variance-covariance for the random effects

$\text{cov}(\mathbf{u},\mathbf{e})=0,$ so that

$$
\begin{aligned}
V(\mathbf{y}) &= V(\mathbf{Xb}+\mathbf{Zu}+\mathbf{e}) \\
&= V(\mathbf{Zu}+\mathbf{e}) \\
&= \mathbf{Z}V(\mathbf{u})\mathbf{Z}'+V(\mathbf{e})+\mathbf{Z}Cov(\mathbf{u},\mathbf{e})+Cov(\mathbf{e},\mathbf{u})\mathbf{Z}' \\
&= \mathbf{ZGZ}'+\mathbf{R}
\end{aligned}
$$

$$Cov(\mathbf{y},\mathbf{u})=\mathbf{ZG}$$

$$Cov(\mathbf{y},\mathbf{e})=\mathbf{R}$$

If we call

$$\mathbf{V}=\mathbf{ZGZ}'+\mathbf{R}$$

$\mathbf{y} \sim \text{N}(\mathbf{Xb}, \mathbf{V})$;
$\mathbf{u} \sim \text{N}(0, \mathbf{G})$;
$\mathbf{e} \sim \text{N}(0, \mathbf{R})$

A trivial example – Daughters Lactation Yield

Herd	Sire	Yield
1	ZA	110
1	AD	100
2	BB	110
2	AD	100
2	AD	100
3	CC	110
3	CC	110
3	AD	100
3	AD	100

$$110 = herd_1 + sire_{ZA} + e$$
$$100 = herd_1 + sire_{AD} + e$$
$$110 = herd_2 + sire_{BB} + e$$
$$100 = herd_2 + sire_{DD} + e$$
$$100 = herd_2 + sire_{DD} + e$$
$$110 = herd_3 + sire_{CC} + e$$
$$110 = herd_3 + sire_{CC} + e$$
$$100 = herd_3 + sire_{AD} + e$$
$$100 = herd_3 + sire_{AD} + e$$

$$\mathbf{y} = \begin{bmatrix} 110 \\ 100 \\ 110 \\ 100 \\ 100 \\ 110 \\ 110 \\ 100 \\ 100 \end{bmatrix} \mathbf{b} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{u} = \begin{bmatrix} s_{ZA} \\ s_{BB} \\ s_{CC} \\ s_{AD} \end{bmatrix} \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \end{bmatrix}$$

$$\mathbf{y = Xb + Zu + e}$$

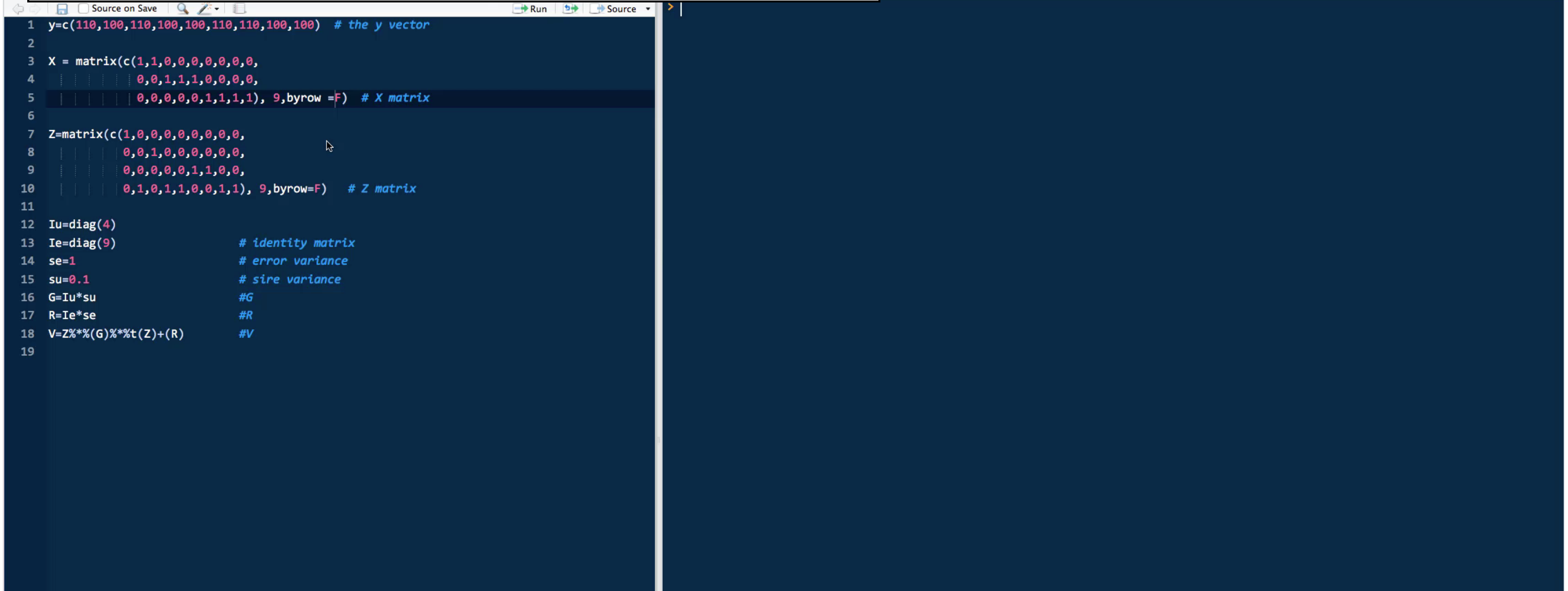
$$V_u = \mathbf{G} = \mathbf{I}\sigma_u^2$$
$$V_e = \mathbf{R} = \mathbf{I}\sigma_e^2$$
$$V_y = \mathbf{Z'GZ + R}$$

y ~N(Xb, V);
u ~N(0, G);
e~ N(0, R)

A trivial example – Daughters Lactation Yield

```
y=c(110,100,110,100,100,110,110,100,100)  # the y vector
X=matrix(c(1,1,0,0,0,0,0,0,0,
           0,0,1,1,1,0,0,0,0,
           0,0,0,0,0,1,1,1,1), 9,byrow=F)  # X matrix
Z=matrix(c(1,0,0,0,0,0,0,0,0,
           0,0,1,0,0,0,0,0,0,
           0,0,0,0,0,1,1,0,0,
           0,1,0,1,1,0,0,1,1), 9,byrow=F)  # Z matrix

Iu=diag(4)
Ie=diag(9)                                 # identity matrix
se=1                                       # error variance
su=0.1                                     # sire variance
G=Iu*su                                    #G
R=Ie*se                                    #R
V=Z%*%(G)%*%t(Z)+(R)                       #V
```



```
y=c(110,100,110,100,100,110,110,100,100)  # the y vector

X = matrix(c(1,1,0,0,0,0,0,0,0,
             0,0,1,1,1,0,0,0,0,
             0,0,0,0,0,1,1,1,1), 9,byrow =F)  # X matrix

Z=matrix(c(1,0,0,0,0,0,0,0,0,
           0,0,1,0,0,0,0,0,0,
           0,0,0,0,0,1,1,0,0,
           0,1,0,1,1,0,0,1,1), 9,byrow=F)   # Z matrix

Iu=diag(4)
Ie=diag(9)                  # identity matrix
se=1                        # error variance
su=0.1                      # sire variance
G=Iu*su                     #G
R=Ie*se                     #R
V=Z%*%(G)%*%t(Z)+(R)        #V
```

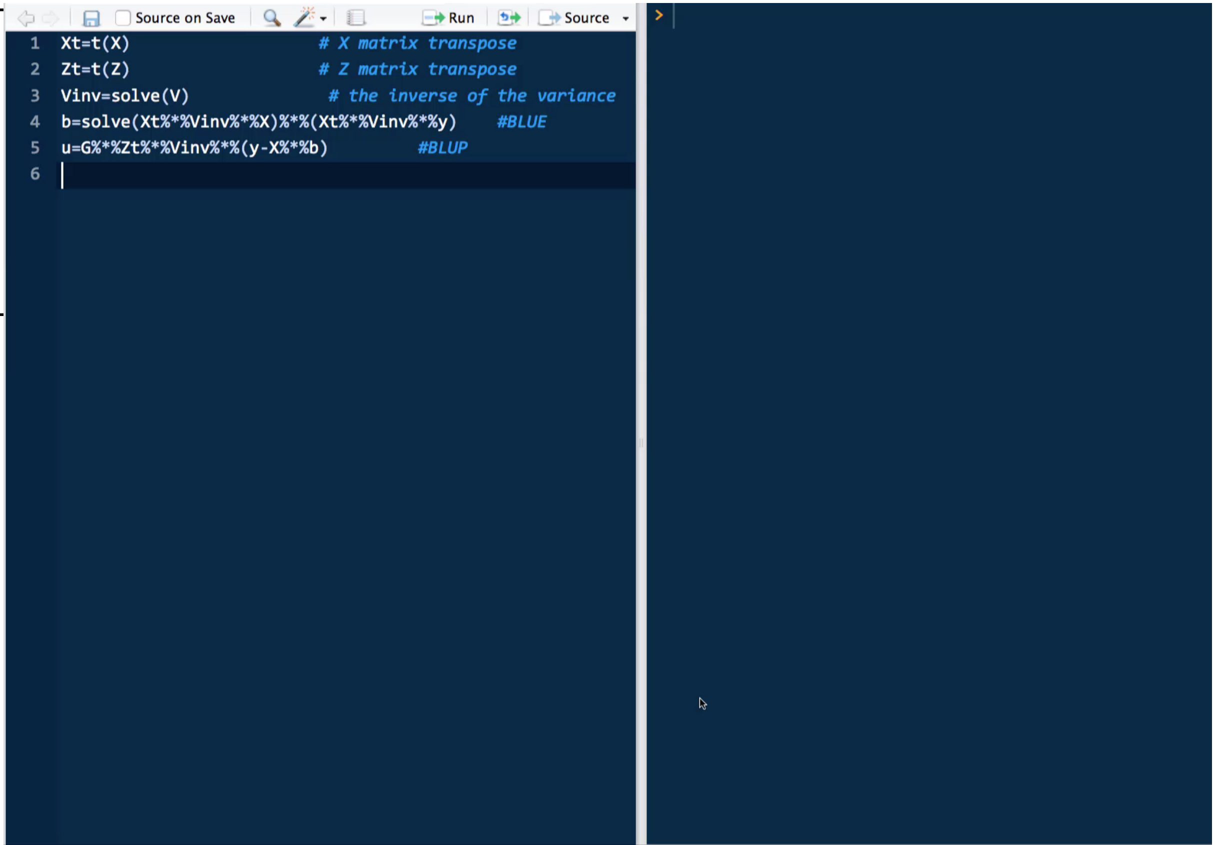
Solving the model

consider variance components as a fixed quantity estimated a priori.

$$\hat{b} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$\hat{u} = GZ'V^{-1}(y - X\hat{b})$$

```
Xt=t(X)
transpose
Zt=t(Z)
transpose
Vinv=solve(V)
variance
b=solve(Xt%*%Vinv%*%X)%*%(Xt%*%Vinv%*%y)
u=G%*%Zt%*%Vinv%*%(y-X%*%b)
```



```
Xt=t(X)                        # X matrix transpose
Zt=t(Z)                        # Z matrix transpose
Vinv=solve(V)                   # the inverse of the variance
b=solve(Xt%*%Vinv%*%X)%*%(Xt%*%Vinv%*%y)     #BLUE
u=G%*%Zt%*%Vinv%*%(y-X%*%b)           #BLUP
```

$$\hat{b} = (105.64, 104.28, 105.46)'$$

$$\hat{u} = (0.40, 0.52, 0.76, -1.67)'$$

The mixed model equations

For the general mixed linear model described above, a particular set of equation can be used in finding the solutions of each effect. These are the Henderson's mixed model equations and were developed for animal breeding by C.R. Henderson (Henderson, 1949).

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{Z'R^{-1}X} \\ \mathbf{X'R^{-1}Z} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

We usually call this the Right Hand Side (RHS)

We usually call this the Left Hand Side (LHS)

The mixed model equations

If we assume that residual variance is IID (identical and independent for all observations), the R matrix can be factored out. In most of the applications we will see from now on the following form of the equations will be more convenient and therefore used:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{Z'X} \\ \mathbf{X'Z} & \mathbf{Z'Z}+\mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$\alpha = \frac{\sigma_e^2}{\sigma_u^2}$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

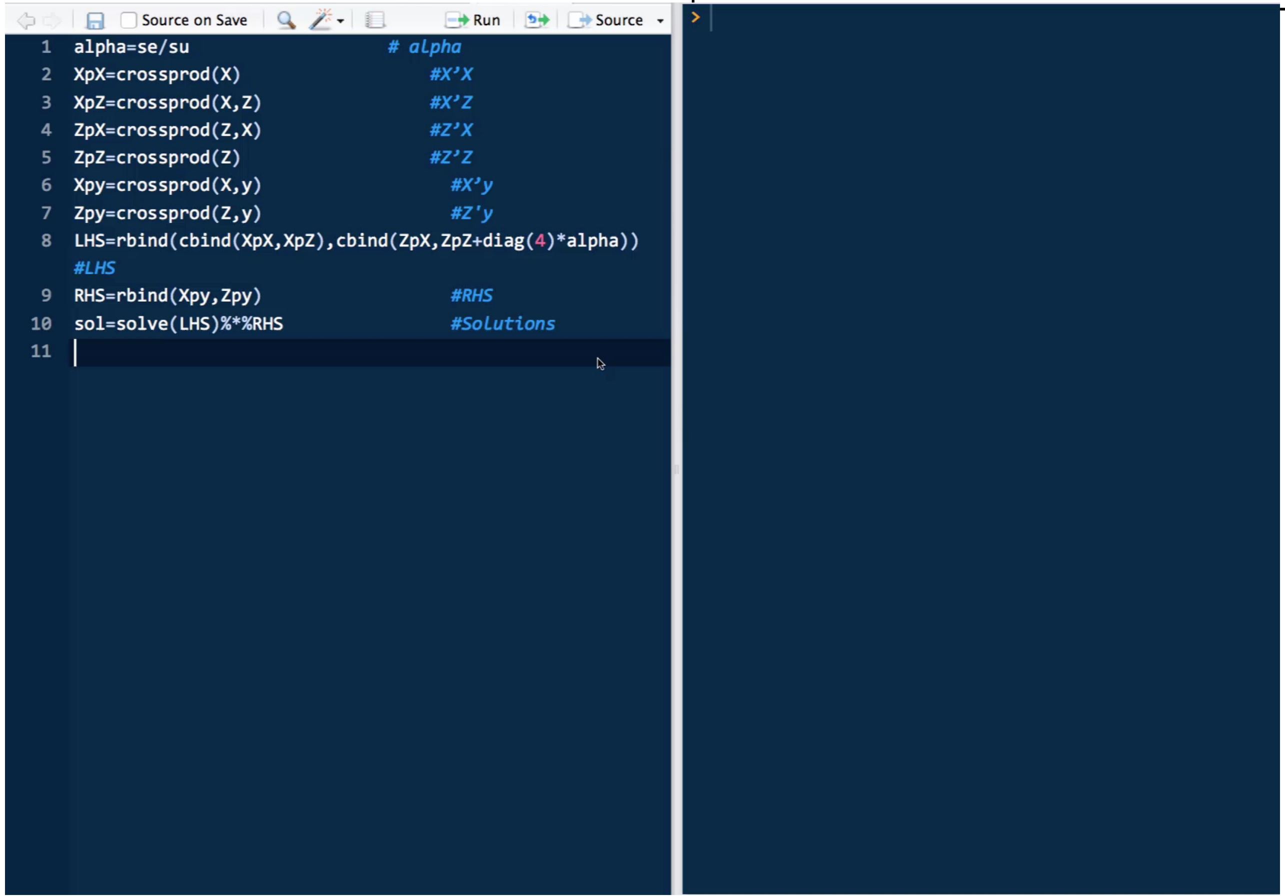
BLUE fixed effects

$$\begin{bmatrix} \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\alpha \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

BLUP random effects

The mixed model equations

```
alpha=se/su                                              # alpha
XpX=crossprod(X)                                         #X'X
XpZ=crossprod(X,Z)                                       #X'Z
ZpX=crossprod(Z,X)                                       #Z'X
ZpZ=crossprod(Z)                                         #Z'Z
Xpy=crossprod(X,y)                                         #X'y
Zpy=crossprod(Z,y)                                         #Z'y
LHS=rbind(cbind(XpX,XpZ),cbind(ZpX,ZpZ+diag(4)*alpha)) #LHS
RHS=rbind(Xpy,Zpy)                                         #RHS
sol=solve(LHS)%*%RHS                                     #Solutions
```



```
alpha=se/su                         # alpha
XpX=crossprod(X)                        #X'X
XpZ=crossprod(X,Z)                      #X'Z
ZpX=crossprod(Z,X)                      #Z'X
ZpZ=crossprod(Z)                        #Z'Z
Xpy=crossprod(X,y)                        #X'y
Zpy=crossprod(Z,y)                        #Z'y
LHS=rbind(cbind(XpX,XpZ),cbind(ZpX,ZpZ+diag(4)*alpha))
#LHS
RHS=rbind(Xpy,Zpy)                        #RHS
sol=solve(LHS)%*%RHS                      #Solutions
```

$$\hat{b}=(105.64,104.28,105.46)'$$

$$\hat{u}=(0.40,0.52,0.76,-1.67)'$$

BLUE and BLUP

While we will not embark in a description of the statistical properties of BLUP and BLUE a reminder of why we like them so much is in order.

They maximize correlations between true values and predicted values (they are BEST) among the (LINEAR) functions of the observations.

The estimates of the fixed effects are (UNBIASED) and the mean of the true (unknown) random effects is equal to the mean of the predicted random effects.

Estimability in models with multiple fixed effects

When there are multiple effects in the model it is often impossible to obtain unique BLUE for each level of the fixed effects.

Herd	Sire	Yield
1	ZA	110
1	AD	100
2	BB	110
2	AD	100
2	AD	100
3	CC	110
3	CC	110
3	AD	100
3	AD	100

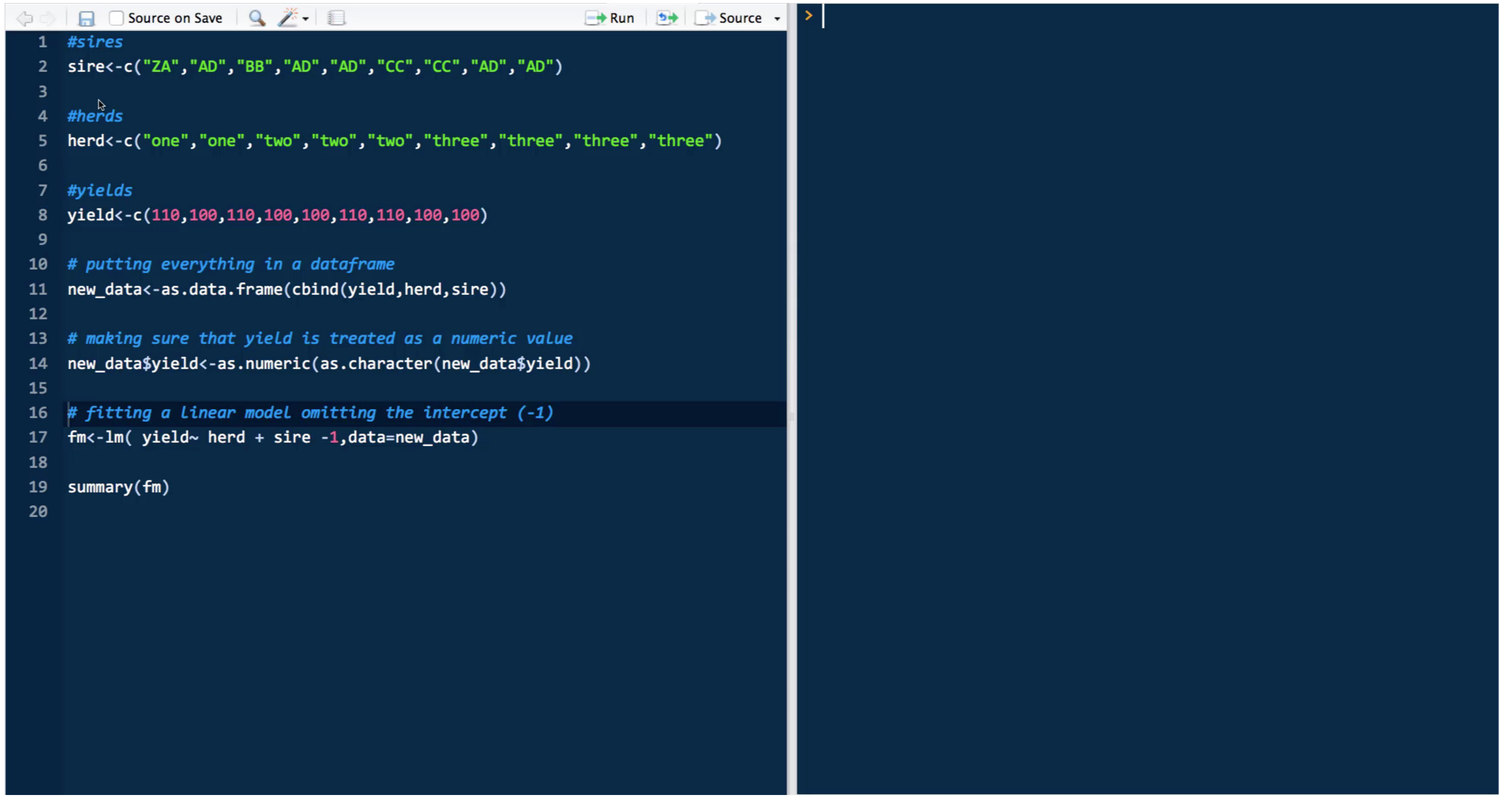
The fourth column (in red) is equal to the difference of the other columns.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & {\color{red}1} & 0 & 0 & 0 \\ 1 & 0 & 0 & {\color{red}0} & 0 & 0 & 1 \\ 0 & 1 & 0 & {\color{red}0} & 1 & 0 & 0 \\ 0 & 1 & 0 & {\color{red}0} & 0 & 0 & 1 \\ 0 & 1 & 0 & {\color{red}0} & 0 & 0 & 1 \\ 0 & 0 & 1 & {\color{red}0} & 0 & 1 & 0 \\ 0 & 0 & 1 & {\color{red}0} & 0 & 1 & 0 \\ 0 & 0 & 1 & {\color{red}0} & 0 & 0 & 1 \\ 0 & 0 & 1 & {\color{red}0} & 0 & 0 & 1 \end{bmatrix}$$

X'X is not full rank since its dimension is 7x7 yet there are only 6 independent rows and columns. In this case a unique inverse of the coefficient matrix (**X'X**) does not exist. Therefore we cannot obtain the BLUE estimates for herd and sire.

Linear functions of the solutions are still estimable.

```
sire<-c("ZA","AD","BB","AD","AD","CC","CC","AD","AD") #sires
herd<-c("one","one","two","two","two","three","three","three","three") #herds
yield<-c(110,100,110,100,100,110,110,100,100) #yields
new_data<-as.data.frame(cbind(yield,herd,sire)) # putting everything in a dataframe
new_data$yield<-as.numeric(as.character(new_data$yield)) # yield as numeric
fm<-lm( yield~ herd + sire -1,data=new_data) # omitting the intercept (-1)
summary(fm)
```



```
#sires
sire<-c("ZA","AD","BB","AD","AD","CC","CC","AD","AD")

#herds
herd<-c("one","one","two","two","two","three","three","three","three")

#yields
yield<-c(110,100,110,100,100,110,110,100,100)

# putting everything in a dataframe
new_data<-as.data.frame(cbind(yield,herd,sire))

# making sure that yield is treated as a numeric value
new_data$yield<-as.numeric(as.character(new_data$yield))

# fitting a linear model omitting the intercept (-1)
fm<-lm( yield~ herd + sire -1,data=new_data)

summary(fm)
```

Set to 0 the first level of the sire fixed effect (sire AD in this case). Model was reparametrized to be full rank the solutions presented are estimable function of the (unknown) BLUEs

$$Sire_{AD} - Sire_{BB}$$

$$Sire_{AD} - Sire_{CC}$$

$$Sire_{AD} - Sire_{ZA}$$

Standard errors and accuracy of the estimates

Accuracy refers to the correlation between true and predicted random effects. Calculation of these values requires knowledge of the inverse elements of the mixed model equations.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{Z'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\alpha \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix}$$

LHS^{-1}

$$PEV = V(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}^{22}\sigma_e^2$$

$$PEV_i = (d_i\sigma_e^2)$$

Diagonal element of C^{22}

Standard errors and accuracy of the estimates

```
round(solve(LHS), digit=3)
        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,]  0.547  0.030  0.024 -0.050 -0.003 -0.004 -0.044
[2,]  0.030  0.383  0.031 -0.003 -0.035 -0.005 -0.057
[3,]  0.024  0.031  0.297 -0.002 -0.003 -0.050 -0.045
[4,] -0.050 -0.003 -0.002  0.095  0.000  0.000  0.004
[5,] -0.003 -0.035 -0.003  0.000  0.094  0.000  0.005
[6,] -0.004 -0.005 -0.050  0.000  0.000  0.092  0.008
[7,] -0.044 -0.057 -0.045  0.004  0.005  0.008  0.083
```

Prediction error variance

$$PEV = \begin{bmatrix} 0.095 \\ 0.094 \\ 0.092 \\ 0.083 \end{bmatrix}$$

As an aside, the sum of each row in C22 is equal to the variance of true sire effect values, which in this case we defined as 0.1 (rounding error causes these example values to be a little bit off):

```
0.095 + 0.000 + 0.000 + 0.004 ~ 0.099
0.000 + 0.094 + 0.000 + 0.005 ~ 0.099
0.000 + 0.000 + 0.092 + 0.008 ~ 0.099
0.004 + 0.005 + 0.008 + 0.083 ~ 0.099
```

Reliability

$$REL = V(true\_values) - V(PE)/V(true\_values)$$

Another measure of the accuracy of the BLUPs is reliability, defined as the squared correlation between predicted and true values (reliability ranges from 0 to 1). For each BLUP *i*, we can compute reliability as (V(true values) – PEV$_i$)/V(true values).

Sire	BLUP	SEP	REL
ZA	0.40	0.308	0.05
BB	0.52	0.306	0.06
CC	0.76	0.303	0.08
ZD	-1.67	0.288	0.17

(0.1 – 0.095)/0.1

Mixed models compared to traditional ANOVA

In this section, we focus on hypothesis testing and estimation of an empirical data set to show how these analyses are conducted for different methods and for different software packages. Only a few details of the mathematical machinery involved in the mixed models analysis will be covered here.

A more detailed description of mixed model theory will be covered in later sections of the book. For readers interested in a more formal treatment of the argument details, they can be found in Sorensen and Gianola (3). Details of ANOVA for balanced and unbalanced data can be found in Rawlings et al. (2001) and Milliken and Johnson (2004).

Mixed models compared to traditional ANOVA

Example data (MaizeRILs.csv) were obtained by testing 62 recombinant inbred line (RIL) progeny from the cross between inbred maize lines B73 and Mo17.

RILs were grown in experimental units (plots) of 20 plants each using a randomized complete block design with two replications at each of four locations.

Five plants in each plot were measured for height.

The mean height for each plot is the independent variable used for this experiment.

Some data were missing from the actual data set, these were filled in with simulated data to create a balanced data set for demonstration.

location	rep	block	plot	RIL	pollen	silking	ASI	height
ARC	1	1	1	RIL-53	74	77	3	184.8
ARC	1	1	2	RIL-40	75	75	0	225.2
ARC	1	1	4	RIL-41	74	74	0	174.4
ARC	1	1	5	RIL-28	69	71	2	147.6
ARC	1	1	6	RIL-11	69	71	2	181.6

location	Location of the progeny test
rep	Replication number
block	Block number. There were 2 blocks at each location
plot	Plot number
RIL	Recombinant inbred line ID
pollen	Days to pollen shed
silking	Days to silking
ASI	Dnthesis-silk interval (*silking* – *pollen*)
height	Mean height of five plants in each plot

NC STATE UNIVERSITY

Mixed models compared to traditional ANOVA

The linear model for this experiment is:

$$Y_{ijk} = \mu + L_i + B(L)_{ij} + G_k + GL_{ik} + \varepsilon_{ijk},$$

Where
μ = overall mean,
L_i = effect of location *i*,
$B(L)_{ij}$ = effect of block *j* nested within location *i* (replication effect),
G_k = effect of genotype *k* (RIL effect),
GL_{ik} = effect of interaction between genotype *k* and location *i*,
ε_{ijk} = residual (experimental error) effect of the plot containing genotype *k* in block *j* of location *i*.

We will assume that all effects except the overall mean are random

ANOVA

BALANCED DESIGN

Source	df	Expected mean squares
Location (L)	n_l-1	$\sigma^2_\varepsilon + n_b\sigma^2_{GL} + n_g\sigma^2_{B(L)} + n_b n_g\sigma^2_L$
Rep(Location) (B)	$(n_b$-$1)\ n_l$	$\sigma^2_\varepsilon + n_g\sigma^2_{B(L)}$
RIL (G)	n_g-1	$\sigma^2_\varepsilon + n_b\sigma^2_{GL} + n_b n_l\sigma^2_G$
RIL*Location (GL)	$(n_g$-$1)(n_l$-$1)$	$\sigma^2_\varepsilon + n_b\sigma^2_{GL}$
Error	$(n_g$-$1)(n_b$-$1)n_l$	σ^2_ε
Total	$n_l n_b n_g$ -1	

ANOVA SAS GLM

```
proc glm data = ril.maizeril;
class location rep RIL;
model height = location rep(location) RIL RIL*location;
random location rep(location) RIL RIL*location/test;
lsmeans RIL;
run;
```

Default F-tests for each factor shown here are correct only for the model in which all effects except residuals are fixed.

Default F-tests use the residual error variance as the denominator in all cases.

Since we have assumed that all effects are random, the correct form of the F-test depends on the expected mean squares.

```
The GLM Procedure
Dependent Variable: mean_height
                                         Sum of
Source                DF        Squares    Mean Square  F Value   Pr >
F

Model                251    264462.4916     1053.6354    16.24
<.0001
Error                244     15832.2400       64.8862
Corrected Total      495    280294.7316

R-Square      Coeff Var       Root MSE     mean_height Mean
0.943516       4.423030       8.055199             182.1195

Source                DF      Type I SS    Mean Square  F Value  Pr > F
location               3     84931.3312     28310.4437   436.31  <.0001
rep(location)          4      3594.2244       898.5561    13.85  <.0001
RIL                   61    154937.5322      2539.9595    39.14  <.0001
location*RIL         183     20999.4038       114.7508     1.77  <.0001

Source                DF    Type III SS    Mean Square  F Value  Pr > F
location               3     84931.3312     28310.4437   436.31  <.0001
rep(location)          4      3594.2244       898.5561    13.85  <.0001
RIL                   61    154937.5322      2539.9595    39.14  <.0001
location*RIL         183     20999.4038       114.7508     1.77  <.0001
```

*F = MS(RIL)/MS(location*RIL) = 2539.959/114.7508 = 22.13 with 61 and 183 df.*

```
proc glm;
class location rep RIL;
model mean_height = location rep(location) RIL RIL*location;
random location rep(location) RIL RIL*location/test;
lsmeans RIL;
run;
```

```
Tests of Hypotheses for Random Model Analysis of Variance
Dependent Variable: mean_height

Source            DF    Type III SS    Mean Square  F Value   Pr > F
location           3          84931          28310    29.85   0.0022
Error         4.4543    4224.553277     948.420722
Error: MS(rep(location)) + MS(location*RIL) - MS(Error)


Source             DF    Type III SS    Mean Square  F Value   Pr > F
rep(location)       4    3594.224445     898.556111    13.85   <.0001
location*RIL      183          20999     114.750840     1.77   <.0001
Error: MS(Error) 244          15832      64.886230


Source             DF    Type III SS    Mean Square  F Value   Pr > F
RIL                61         154938    2539.959544    22.13   <.0001
Error             183          20999     114.750840
Error: MS(location*RIL)
```

*F = MS(RIL)/MS(location*RIL) = 2539.959/114.7508 = 22.13 with 61 and 183 df.*

ANOVA SAS GLM

$$\hat{\sigma}_{\varepsilon}^{2} = MS(error) = 64.89$$

$$\hat{\sigma}_{GL}^{2} = \frac{MS(location * RIL) - MS(error)}{n_b} = \frac{114.75 - 64.89}{2} = 24.93$$

$$\hat{\sigma}_{G}^{2} = \frac{MS(RIL) - MS(location * RIL)}{n_l n_b} = \frac{2539.96 - 114.75}{8} = 303.15$$

Var Comp. method of moments → equate the observed mean squares to their expectations and solve for the variance components

Predicted marginal mean value of each RIL using the “lsmeans”

```
Least Squares Means
           mean_height
RIL             LSMEAN
RIL-1       182.100000
RIL-11      182.875000
RIL-12      185.200000
RIL-14      194.250000
RIL-15      195.775000
RIL-16      172.825000
RIL-20      209.750000
RIL-21      165.850000
```

ANOVA R

```r
#ANOVA
m0 <- lm(mean_height~location*RIL + rep:location)
anova(model)
# Polulation means ( LSMEANS )
pma <- popMatrix(mm,effect='RIL') # obtaining L
RILMeans <- popMeans(model, effect='RIL')
```

```
Analysis of Variance Table

Response: mean_height
              Df   Mean Sq   F value    Pr(>F)
location       3   28310.4  436.3090  < 2.2e-16 ***
RIL           61    2540.0   39.1448  < 2.2e-16 ***
location:RIL 183     114.8    1.7685  1.643e-05 ***
location:rep   4     898.6   13.8482  3.408e-10 ***
Residuals    244      64.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Full sample code is available in “example2a.r” file.

Balanced data – mixed models analysis with SAS Proc MIXED

```
proc mixed covtest;
class location rep RIL;
model mean_height = /solution;
random location rep(location) RIL RIL*location/solution;
ods output solutionR = random solutionF = fixed;
```

- No degrees of freedom, sum of squares, mean squares, or F-tests for random terms.
- VarComp estimated
- REML varcomp=Moment method with balanced design

```
                     The Mixed Procedure


              Covariance Parameter Estimates
                             Standard          Z
Cov Parm          Estimate      Error      Value          Pr Z
location            220.66     186.48       1.18        0.1184
rep(location)      13.4463    10.2484       1.31        0.0948
RIL                 303.15    57.5088       5.27        <.0001
location*RIL       24.9323     6.6787       3.73        <.0001
Residual           64.8862     5.8745      11.05        <.0001


          Fit Statistics
-2 Res Log Likelihood         3833.2
AIC (smaller is better)       3843.2
AICC (smaller is better)      3843.3
BIC (smaller is better)       3840.1
```

Balanced data – mixed models analysis with R

```
require(lme4, quiet=T)
mm <-lmer(mean_height ~ 1 + (1|RIL) + (1|location/rep) +
(1|location:RIL))
print(mm)
```

```
Linear mixed model fit by REML
Formula: mean_height ~ 1 + (1 | RIL) + (1 | location/rep) + (1 |
location:RIL)
   Data: rils
  AIC  BIC logLik deviance REMLdev
 3845 3870  -1917     3839    3833

Random effects:
 Groups       Name        Variance Std.Dev.
 location:RIL (Intercept)  24.932   4.9932
 RIL          (Intercept) 303.151  17.4112
 rep:location (Intercept)  13.446   3.6669
 location     (Intercept) 220.661  14.8547
 Residual                  64.886   8.0552
Number of obs: 496, groups: location:RIL, 248; RIL, 62;
rep:location, 8; location, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  182.119      7.869   23.14
```

Hypothesis testing with mixed models

Hypothesis testing for the variance components can be based on the “Z value” obtained by using the “covtest” option in the Proc MIXED statement.

The Z value is the ratio of the variance component to its standard error, which has a Z distribution. This test has low power, particularly for variance components estimated with few degrees of freedom.

Hypothesis testing with higher power can be implemented with the likelihood ratio test.

This test requires one to perform an additional Proc MIXED analysis for each factor to be tested, in which one removes the factor of interest from the model.

The likelihood of this “reduced” model can be compared to the likelihood of the “full” model to form a test of the null hypothesis that the variance component for the dropped term is zero.

If removing the term causes a large decrease in the likelihood of the model, then there is more evidence that the variance component for the term is greater than zero.

Test significance of location main effect by dropping it from model, compute likelihood of reduced model:

```
proc mixed;
class rep RIL;
model mean_height =;
random rep(location) RIL RIL*location;
```

```
              Fit Statistics
-2 Res Log Likelihood            3841.3
```

```
LRT = (-2) ln (likelihood of reduced model / likelihood of full model)
= -2 * ln (likelihood of reduced model) – (-2)*ln (likelihood of full
model)
= 3841.3 – 3833.2 = 8.1
```

LRT has approx. chi-square distribution. DF equal the number of parameters (variance components) that differ between the models. The raw *p*-value should be divided by 2.
H_0: $\sigma^2_{location} = 0$, raw *p*-value of χ^2 with value 8.1 and df of 1 = 0.004. Adjusted *p*-value = 0.002

Model	-2 RLL	LRT	raw p-value	adjusted p-value
Full	3833.2			
No RIL	4092.7	259.5	<0.0001	<0.0001
No location*RIL	3850.4	17.20	<0.0001	<0.0001

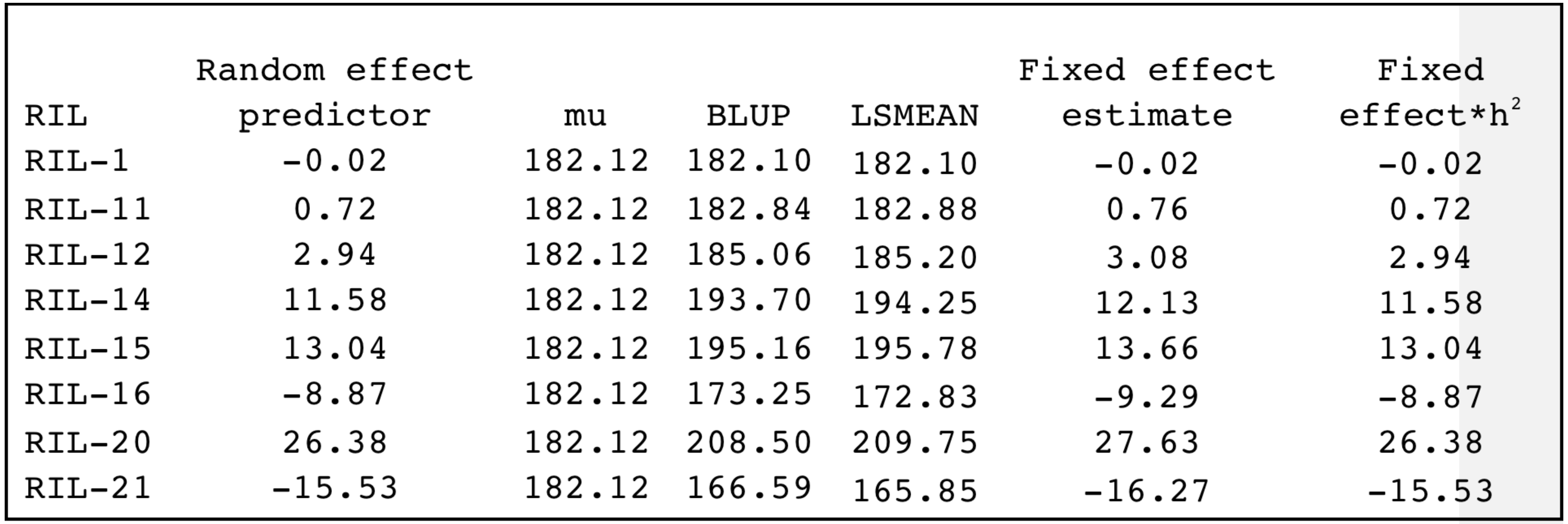
Prediction of random factors with mixed models

To compare the estimation of RIL values from Proc MIXED and Proc GLM, the first difference one will notice is that an error message will result if one includes the statement “lsmeans RIL” as part of the Proc MIXED analysis.

In practice, we obtain the RIL effect predictions by requesting the solutions for the random effects in the model using the “/solution” option on the random statement in Proc MIXED.

We can then construct best linear unbiased predictors (BLUP) in this case by simply adding the estimated overall mean effect (μ, obtained with the “/solution” option on the model statement) to each RIL effect prediction

```
                  Solution for Fixed Effects
                                 Standard
Effect              Estimate         Error      DF    t Value      Pr >
|t|
Intercept             182.12        7.8719       3      23.14
0.0002
```



RIL	Random effect predictor	mu	BLUP	LSMEAN	Fixed effect estimate	Fixed effect*h^2
RIL-1	-0.02	182.12	182.10	182.10	-0.02	-0.02
RIL-11	0.72	182.12	182.84	182.88	0.76	0.72
RIL-12	2.94	182.12	185.06	185.20	3.08	2.94
RIL-14	11.58	182.12	193.70	194.25	12.13	11.58
RIL-15	13.04	182.12	195.16	195.78	13.66	13.04
RIL-16	-8.87	182.12	173.25	172.83	-9.29	-8.87
RIL-20	26.38	182.12	208.50	209.75	27.63	26.38
RIL-21	-15.53	182.12	166.59	165.85	-16.27	-15.53

$$\hat{h}^2 = \frac{\hat{\sigma}^2_{RIL}}{\hat{\sigma}^2_{RIL} + \frac{\hat{\sigma}^2_{RIL*location}}{4} + \frac{\hat{\sigma}^2_{\varepsilon}}{8}} = 0.954$$

Unbalanced data - ANOVA with SAS Proc GLM

```
The GLM Procedure
Dependent Variable: mean_height
                                    Sum of
Source                  DF         Squares   Mean Square  F Value  Pr > F
Model                  248     251616.8381     1014.5840    15.06  <.0001
Error                  225      15157.9644       67.3687
Corrected Total        473     266774.8025

R-Square     Coeff Var      Root MSE    mean_height Mean
0.943181      4.509687      8.207846            182.0048

Source                  DF     Type I SS   Mean Square  F Value  Pr > F
location                 3    79345.6274    26448.5425   392.59  <.0001
rep(location)            4     3693.2364      923.3091    13.71  <.0001
RIL                     61   150287.3376     2463.7268    36.57  <.0001
location*RIL           180    18290.6367      101.6146     1.51  0.0018

Source                  DF   Type III SS   Mean Square  F Value  Pr > F
location                 3    77171.8286    25723.9429   381.84  <.0001
rep(location)            4     3677.7950      919.4488    13.65  <.0001
RIL                     61   149644.4648     2453.1879    36.41  <.0001
location*RIL           180    18290.6367      101.6146     1.51  0.0018
```

Unbalanced data - ANOVA with SAS Proc GLM

- First, notice that the degrees of freedom for RIL are still 61 but the degrees of freedom for location*RIL are now 180 instead of 183.
- The reason for this is that we have no data on two of the location*RIL interactions involving RIL-5 (because we have no data on this RIL from two locations) and one of the location*RIL interactions involving RIL-51 (as it is missing in one location).
- Second, note that now the Type I and Type III sums of squares (SS) and mean square (MS) results are different from each other in this case. This occurs because the
  - Type I statistics are computed by fitting the effects in the order given in the model and computing the sums of squares accounting for each term sequentially, whereas the
  - Type III statistics are computed by calculating the sums of squares attributable to each term after accounting for all other terms in the model.
- In the case of balanced data, all of the model terms are orthogonal to each other such that the order of fitting factors affects how much variation they are associated with.
- In contrast, with unbalanced data, the different factors can become correlated and the variation associated with any one term may also be partly associated with a different term, such that the order of fitting terms affects the sums of squares for the term.
- Because of this, Type III statistics are preferred since they indicate the amount of variation attributable to each factor after accounting for the other factors in the model

As a result, the sum of Type III statistics will be less than the total sums of squares for the model: in this example the sum of the Type III SS = 248784.7251, whereas the total SS for the model is 266774.8025.

Unbalanced data - ANOVA with SAS Proc GLM

- Variance components can be estimated by the method of moments from ANOVA Type III MS, but two complications arise in the case of unbalanced data:
    - First, such estimates are reasonable estimates if the data are not too badly balanced, but the statistical properties of such estimators are unknown, so it can be difficult to know how reliable they are for a given data set.
    - Second, the expected mean squares shown for the balanced data set above are not correct for the unbalanced data case, as the coefficients on the variance components are affected by the data structure.
    - The computation of the coefficients can be horribly complex (see Rawlings and Messy Data books), but we can get the coefficients using SAS Proc GLM with the random statement, resulting in this output:

```
Source          Type III Expected Mean Square
location        Var(Error) + 1.8769 Var(location*RIL) + 57.243
                Var(rep(location)) + 114.49 Var(location)

rep(location)   Var(Error) + 57.25 Var(rep(location))
RIL             Var(Error) + 1.9149 Var(location*RIL) +
7.6105Var(RIL)
location*RIL    Var(Error) + 1.9348 Var(location*RIL)
```

$$F_{RIL} = \frac{MS(RIL)}{MS(location * RIL)} = \frac{\sigma_\varepsilon^2 + 2\sigma_{location*RIL}^2 + 8\sigma_{RIL}^2}{\sigma_\varepsilon^2 + 2\sigma_{location*RIL}^2} = 1 + \frac{8\sigma_{RIL}^2}{\sigma_\varepsilon^2 + 2\sigma_{location*RIL}^2}$$

WITH BALANCE

$$F_{RIL} = \frac{MS(RIL)}{MS(location * RIL)} = \frac{\sigma_\varepsilon^2 + 1.9149\sigma_{location*RIL}^2 + 7.6105\sigma_{RIL}^2}{\sigma_\varepsilon^2 + 1.9348\sigma_{location*RIL}^2}$$

WITH NO BALANCE

As this expectation does not equal 1 when the null hypothesis is true (RIL variance component = 0), the F-test is not correct. Instead, more complicated forms of the F-test are required, and Proc GLM computes these forms when the random statement is given:

```
The GLM Procedure
Tests of Hypotheses for Random Model Analysis of Variance
Dependent Variable: mean_height

Source               DF    Type III SS    Mean Square  F Value   Pr > F
location              3          77172          25724    27.00   0.0031

Error            4.2929    4089.266951     952.566239
Error: 0.9999*MS(rep(location)) + 0.9701*MS(location*RIL)
- 0.9699*MS(Error)

Source                  DF    Type III SS    Mean Square  F Value   Pr > F
rep(location)            4    3677.795036     919.448759    13.65   <.0001
location*RIL           180          18291     101.614649     1.51   0.0018

Error: MS(Error)       225          15158      67.368731

Source                  DF    Type III SS    Mean Square  F Value   Pr > F
RIL                     61         149644    2453.187948    24.23   <.0001

Error               182.48          18479     101.262345
Error: 0.9897*MS(location*RIL) + 0.0103*MS(Error)
```

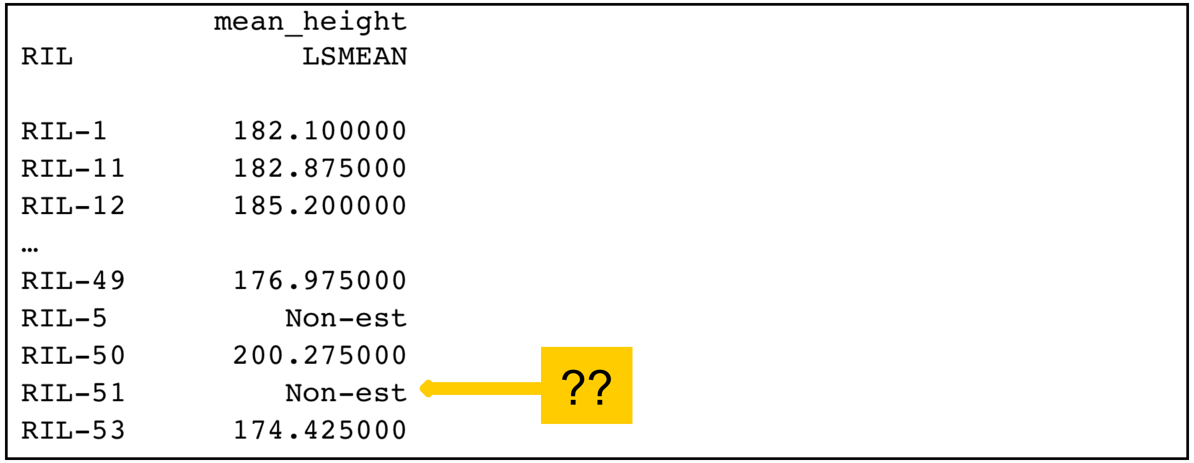
Unbalanced data - ANOVA with SAS Proc GLM

$$\hat{\sigma}_{\varepsilon}^{2} = MS(error) = 67.37$$

$$\hat{\sigma}_{GL}^{2} = \frac{MS(location * RIL) - MS(error)}{1.9348} = \frac{101.61 - 67.37}{1.9348} = 17.7$$

$$\hat{\sigma}_{G}^{2} = \frac{MS(RIL) - \ MS(location * RIL)}{8} = \frac{2539.96 - 114.75}{8} = 303.15$$

LSmeans



```
               mean_height
RIL                 LSMEAN

RIL-1           182.100000
RIL-11          182.875000
RIL-12          185.200000
…
RIL-49          176.975000
RIL-5              Non-est
RIL-50          200.275000
RIL-51             Non-est
RIL-53          174.425000
```

??

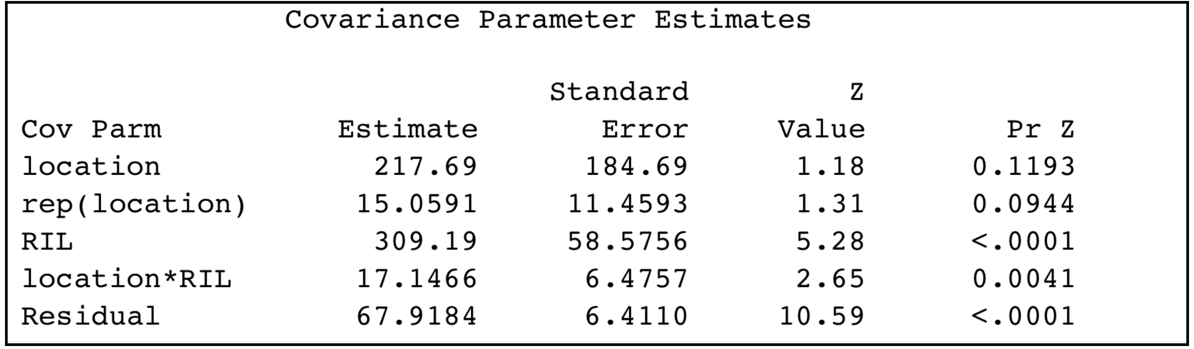
Unbalanced data - ANOVA with SAS Proc GLM

$$(Y_{..k}) = \mu + \bar{L}_{.} + \overline{B(L)}_{..} + G_k + \overline{GL}_{.k}$$

In the fixed model, the interaction of a genotype-by-location interaction effect is non-estimable if there are no data on that combination of genotype and location. Then if some interaction effects included in the LSmean equation are non-estimable, the whole LSmean is non-estimable

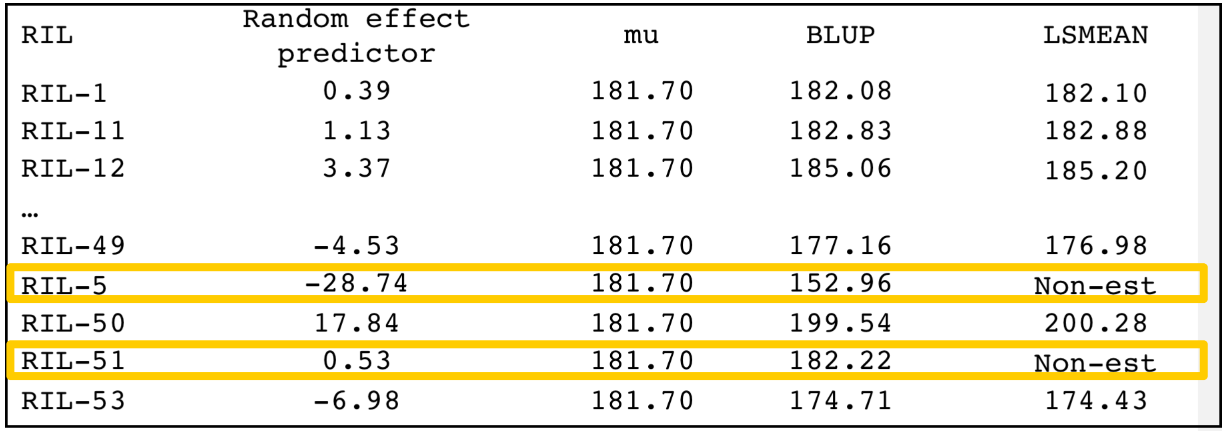
Unbalanced data – mixed models analysis

Same model with proc mixed



```
                    Covariance Parameter Estimates


                                    Standard         Z
Cov Parm            Estimate           Error     Value         Pr Z
location              217.69          184.69      1.18       0.1193
rep(location)        15.0591         11.4593      1.31       0.0944
RIL                   309.19         58.5756      5.28       <.0001
location*RIL         17.1466          6.4757      2.65       0.0041
Residual             67.9184          6.4110     10.59       <.0001
```



RIL	Random effect predictor	mu	BLUP	LSMEAN
RIL-1	0.39	181.70	182.08	182.10
RIL-11	1.13	181.70	182.83	182.88
RIL-12	3.37	181.70	185.06	185.20
…				
RIL-49	-4.53	181.70	177.16	176.98
RIL-5	-28.74	181.70	152.96	Non-est
RIL-50	17.84	181.70	199.54	200.28
RIL-51	0.53	181.70	182.22	Non-est
RIL-53	-6.98	181.70	174.71	174.43

LSmean (BLUE):

BLUP: $\hat{Y}_{..k} = \mu + G_k$