Introduction to Mixed Models

Introduction to Mixed Models

Linear models are the most common type of statistical models used in animal breeding to predict breeding values based on phenotypic observations. Linear models form the basis of Best Linear Unbiased Prediction. A linear model provides a machinery to correct breeding values for systematic environmental effects (usually termed as fixed effects). BLUP estimation of breeding values is based on a mixed model, which is a linear model containing fixed effects as well as random effects (usually the additive genetic values).

Introduction

Why are linear models important?

Data sets in animal breeding are generally used to estimate breeding values and/or genetic parameters. Taking the example of breeding values, different information sources are used to obtain the most precise estimate of an animal's genetic ability. This information consists of measured phenotypes that are influenced not only by the animals' genes, but also by many other environmental effects. A simple 'solution' might be that we take the different measurements as a deviation of a comparable mean. This could be a population mean or, if animals perform in different years and/or different herds, the mean of all animals in that year and/or herd. Such deviation should be free of those environmental effects. Problem with this simple approach are

- Different herds use different sires and their means are not only determined by environment.
- We need to take into account how much information we have to estimate means. An estimate of a herd mean based on 5 animals is less accurate than one with 100 animals.

The main practical advantage if a linear model is that it can appropriate account for all effects that influence a measurement. This is particularly useful when the data is unbalanced, which is nearly always the case in field data, and often also in

experimental data relating to animals. The following example illustrates why a simple approach will not work.

Table 1: Example data to illustrate analysis of unbalanced data

Cow	Breed	Feeding regime	Weight (kg)
1	Angus	intensive	494
2	Angus	intensive	556
3	Angus	extensive	542
4	Hereford	extensive	473
5	Hereford	intensive	632
6	Hereford	extensive	544

In the example, the mean of Angus cows is equal to 530.7 kg and the mean of Hereford cattle is 549.7 kg. Hence, the breed difference from this data could be estimated to be equal to 19 kg. However, we see that the Angus cattle were relatively more fed on an intensive feed. Therefore, the earlier estimate of 19 kg for breed differences is biased by unequal feeding regimes. We would need to know the effect of feeding regime and correct for this. However, the difference between intensive and extensive feeding is also affected by the unequal representation of breeds. A linear model will exactly spell out which effects are affecting which observation and the different effects (such as breed and feeding regime) are estimated simultaneously and during this process they are corrected for each other.

Therefore, a very important reason for using linear models is to account appropriately for unbalancedness in data. Linear models can be advanced and accommodate different effects, covariances between different effects, different types of distributions etc.

The introduction of linear models in animal breeding took place halfway the 20th century, and was mostly related to evaluation of dairy bulls. Differences in the average production of herd mates are caused by differences in environment as well as by differences in genetic level. To obtain unbiased estimated breeding values, effects of sires and effects of herds have to be estimated simultaneously. To achieve this 'mixed models' are used in which fixed effects and breeding values (indicated as 'random effects') will be estimated jointly. This procedure is called "BLUP", and was developed by C.R. Henderson (1949,1973). BLUP stands for Best Linear Unbiased

Prediction, which describes the statistical properties of the estimated breeding values obtained using this method.

In the next chapter, we will elaborate on the difference between the estimation of fixed effects and the prediction of random or stochastic effects (breeding values). We will pursue with presenting mixed models. Using examples, we will indicate how to set up equations, which principles are important and how breeding values are predicted.

Linear Models

Linear models are commonly used to describe and analyse data in the biological sciences. The model needs to represent the sampling nature of the data.
The data vector contains measurements on experimental units. The observations are random variables that follow a multivariate distribution. The model usually consists of factors. These are variables, either discrete or continuous, which have an effect on the observed data. Different model factors are:

- Discrete factors or class variables such as sex, year, herd
- Continuous factors or covariables such as age

Some factors are of special interest to the researcher but other factors have to be included in the model simply because they explain a significant part of the variation in the data and reduce the residual (unexplained) variation. Such factors are often called ‘nuisance variables’.

Fixed and random effects

Another distinction that is often used is that between fixed and random effects. The statistical world is somewhat divided her in more traditional ‘frequentists’ that make this distinction and Bayesians’ that find this distinction artificial and accommodate the properties of different factors in their model specification. However, it is still useful to try to define the difference between fixed and random effects, and acknowledge this dispute.

Fixed Effects

- Effects for which the defined classes comprise all the possible levels of interest, eg. sex, age, breed, contemporary group. Effects can be considered as fixed when the number of levels are relatively small and is confined to this number after repeated sampling.

Random Effects

- Effects which have levels that are considered to be drawn from an infinite large population of levels. Animal effects are often random. In repeated experiments there maybe other animals drawn from the population.

The distinction is also often determined by the purpose of the experiment. Do we want to know the difference between these specific levels of a factor, or are we interested in how large the differences between levels of a factor might generally be. The effect of management groups could be fixed but arguments for considering them as random could be found just as easily.

Example A growth trial for a number of animals from different age groups used several different diets, locations and handlers.

In this case the number of levels for age, diet, location and handler could all conceptually be the same for an infinite number of sampling events. On the other hand different animals would be needed for each repeated sample as the same growth phase could not be repeated in the same animal. Furthermore inferences might be made about diets or locations in general and in this case these effects might be considered random since these could have been sampled from an infinite number of levels. Therefore animals effects would be considered random while all other effects would generally be fixed.

A checklist that can be used for deciding about fixed or random effects:

i) What are the number of levels?

small	-	fixed
large or near infinite	-	possibly random

ii) Are the levels repeatable?

yes	-	fixed
no	-	random

iii) Are there conceptually and infinite number of such levels?

yes	-	possibly fixed
no	-	possibly random

iv) Are inferences to be made about levels not included in the sampling?

yes	-	possibly random
no	-	possibly fixed

v) Were the levels of the factor determined in a non-random manner?

yes	-	possibly random
no	-	possibly fixed

A linear relationship can generally be found to fit most biological data although some transformation may be required. Thus a linear model can generally be used to describe data. All models contain a set of factors composed of three parts which additively affect the observations or records of data:

i) the equation

ii) expectations and variance covariance matrices of random variables

iii) assumptions, limitations and restrictions

The Equation

The equation of a model defines the factors that will or could have an effect on an observed trait. The general linear model equation in matrix form is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad ...(1)$$

where

y is an $n \times 1$ vector of n observed records

b is a $p \times 1$ vector of p levels of fixed effects

u is a $q \times 1$ vector of q levels of random effects

e is an $n \times 1$ vector of random, residual terms

X is a known *design matrix* of order $n \times p$, which relates the records in **y** to the fixed effects in **b**

Z is a known *design matrix* of order $n \times q$, which relates the records in **y** to the random effects in **u**

Equation (1) is generally termed a *mixed model* as it contains both fixed and random effects. While not specified directly, interactions between fixed effects are fixed, interactions between random effects are random and interactions between fixed and random effects are random. The mixed model can be reduced to become a fixed effect

model by not including **Zu** or a random effects model for which no fixed effects are fitted except the overall mean, i.e. **Xb** = 1μ.

Expectations and Variance Covariance (VCV) Matrices

In general the expectation of **y** is

$$E\begin{pmatrix} y \\ u \\ e \end{pmatrix} = \begin{pmatrix} Xb \\ 0 \\ 0 \end{pmatrix} \qquad ...(2)$$

which is also known as the 1st moment. The 2nd moments describe the variance-covariance structure of y:

$$V\begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \qquad ...(3)$$

where **G** is a dispersion matrix for random effects other than errors and **R** is the dispersion matrix of error terms, for which both are general square matrices assumed to be non-singular and positive definite, with elements that are assumed known.

We usually write

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

Assumptions, Limitations and Restrictions

This part of the model identifies any differences between the operational and ideal models. It may describe the sampling process and to which extend the assumptions that are made can be expected to be true (e.g. about normality, random sampling, uncorrelated error terms, equally distributed error terms, etc).

Estimation Theory

Estimating Fixed Effects

Consider a general model

$$y = Xb + \varepsilon \qquad ...(4)$$

$$\text{with } E(y) = Xb \quad \text{and} \quad var(y) = V = var(\varepsilon) \qquad \ldots(5).$$

We want to estimate fixed effects in b and conduct hypothesis testing about the significance of differences between the different levels of effects. Note that ε is a vector with random effects. They can be caused by several random factors (e.g. animal and residual) and the different levels may be correlated (e.g. due to repeated measurements on the same animals). Hence, var(ε) maybe equal to V = ZGZ'+R.

To find good estimators of the fixed effects parameters for a set of data, trial and error could be used. However the method of least squares, developed by Gauss in 1809 and Markoff in 1900 is commonly used for estimating these parameters of which the theorem states that

> ... under the conditions of the model as described in (4) and (5) the least squares estimators b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

The proof is given in several texts on linear models. Unbiasedness occurs when E(Xβ) = Xb where β is an estimate of b,. However to estimate the value of these estimates, consideration needs to be given to the deviation of y_i from its expected value

$$E(y) = X\boldsymbol{b} \qquad ...(3)$$

and more importantly to the sum of the *N* squared deviations (errors) given as Q where

$$Q = (y- X\beta)'(y-X\beta) \quad ...(4)$$

According to the method of least squares the *best* estimators of β_0 and β_1 are those which minimise Q.

Best - maximises the correlation between true and estimated value of effects by minimising the error variance.

Linear - the factors for which estimates are required are linear functions of the observations.

Unbiased - estimates of fixed effects and estimable functions are such that $E(\beta \mid b) = b$.

Deriving Estimates Using Ordinary Least Squares

The general fixed effects model in matrix for is

$$y = Xb + e \quad ...(5)$$

where *y* is a vector of observations, X is an incidence matrix linking the independent variables to the observations, *b* is a vector of effects to be solved and *e* is a vector of error terms. For ordinary least squares (OLS), error terms are independently and identically distributed random variables with a mean of zero and a variance of σ_e^2 such that var(y) = var(e) = $I_N \sigma_e^2$ where I_N is a dispersion matrix for *n* observations. Given that E(y) = Xb,

$$Q = (y - X\beta)' (y - X\beta)$$

which when differentiated with respect to *b* gives

$$\delta Q / \delta b = -2(X'y + X'Xb).$$

Equating to zero gives

$$X'Xb = X'y$$

which are referred to as the *normal equations* which if the inverse of X'X exists, provides the least square estimator of β:

$$b = (X'X)^{-1} X'y \qquad ...(6)$$

Thus ordinary least squares assumes that all observations are uncorrelated and have a common variance σ_e^2. If estimates are derived when this is not true then they are no longer 'best' since Q is no longer minimised.

Deriving Estimates Using Generalised Least Squares

For ordinary least squares, the criterion (4) weights each observation equally. However σ_e^2 may not be common to all observations. Let var(e) = V where

$$V = \begin{bmatrix} w_1 & & & \\ & w_2 & 0 & \\ & & .. & \\ & 0 & .. & \\ & & & w_n \end{bmatrix} \sigma_e^2$$

and the dispersion matrix is known. In this case V is still be a diagonal matrix, but now not all elements will be the same. For example suppose sires were being measured by their mean progeny merit. In this case the diagonal elements of V could be weighted by the number of each progeny group. Estimates obtained via this method are generally known as *weighted least squares* (*WLS*).

Alternatively V might be non-diagonal and contain variance components such that

$$V = \begin{bmatrix} v_1 & & & \\ & v_2 & ij & \\ & & .. & \\ & ij & & .. \\ & & & v_n \end{bmatrix}$$

where v_i is the variance of the *i*th observation and *ij* are off diagonal elements and are the covariances between them. An example case would be for observations on groups of half sibs such that there would be covariances between measurements. In most genetic models there is a second random effect (besides error) and there are covariances among the random terms (e.g. due to genetic relationships). Therefore V is generally not diagonal in genetic analysis. This case is conventionally known as *generalised least squares* (*GLS*) where OLS and WLS are merely special cases of GLS. The generalised least squares criterion for simple linear regression is

$$Q_G = Q = (y- X\beta)'V^{-1}(y-X\beta)$$

Minimising Q_G with respect to β_0 and β_1 leads to the appropriate normal equations of

$$(X'V^{-1}X)\,\beta = X'V^{-1}Y$$

Determining a generalised inverse for $X'V^{-1}X$ gives the least square estimates as

$$\beta = (X'V^{-1}X)^{-}\, X'V^{-1}Y \qquad ...(7)$$

which is a general equation for any fixed effects model.

Estimability

Because a generalized inverse of $X'V^{-1}X$ is used there are a large (infinite) number of possible solutions to b. However, any solution vector can be used to compute *estimable* functions of b. An estimable function has the same numeric value, i.e. is unique, for any of the possible solution vectors. The following functions are always estimable:

- Any linear function of y is estimable
- Any linear function of E(y) is estimable
- K'b is estimable if K' = TX for some T, i.e. T is a linear combination of rows in X.
- Q'b is estimable if $Q'(X'V^{-1}X)^{-}X'V^{-1}X = Q'$

Example:

$$\begin{bmatrix} 6 & 4 & 2 \\ 4 & 4 & 0 \\ 4 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 120 \\ 82 \\ 38 \end{bmatrix}$$

has many possible solutions, e.g. β' = [0 20.5 19] or [20 +0.5 -1]. (Verify this)

However, the function $\mu + \alpha_1$ *is* equal to 20.5 for all possible solutions. Also the difference $\alpha_1 - \alpha_2$ is always equal to 1.5. Only estimable functions have a meaning in a statistical analysis because they are unique.

Statistical packages usually give a set of solutions that is based on a constraint. Constraints enforce unique solutions for b, but because the constraints are arbitrary, the solutions are arbitrary as well. Constraints can be enforced by manipulation the X matrix such that it becomes non-singular, i.e. linear combinations of the columns should not be able to result in another linear combination of columns. The following example illustrates estimability and uniqueness of solutions:

Example dataset 2

Year of birth	Sex	weight
1990	Male	354
1990	Female	251
1991	Male	327
1991	Female	328
1991	Male	301
1991	Female	270
1992	Male	330

First consider to fit year of birth:

$y = \mu + year_i + e_{ij}$ *or* y = Xb+e

then the solution for b can be obtained as

$$\beta = (X'X)^{-1} X'y$$

and the matrices look like

$$X'X = \begin{pmatrix} 7 & 2 & 4 & 1 \\ 2 & 2 & 0 & 0 \\ 4 & 0 & 4 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \text{ and } \qquad X'y = \begin{pmatrix} 2161 \\ 605 \\ 1226 \\ 330 \end{pmatrix}$$

and the X'X matrix contains the number of observations and X'Y contains the sum of all the observations. "Dividing' X'Y by X'X gives therefore the average per class.

A complication in this example is that the columns of X add up to each other. This is always the case if we have more than one fixed effect. If the columns add up (i.e. X is singular), also X'X is singular, and can not be inverted. A practical explanation is that we *want* to estimate 4 parameters (a general mean and three year effects), but in our data we have only three year means, so we *can* only estimate three parameters as we have only three independent means. We can find solutions by setting a restriction:

1) put the general mean to zero
2) put one of the years to zero
3) put the sum of the year effects to zero

NB: The option you choose is arbitrary, it does effect the estimates, but not the relevant comparisons, in this case, it does not affect the estimate of the year difference!

The second option is the easiest. It means that you leave the equation for the year that you give a zero solution out of the equations, and the general mean will be in fact the estimate of the mean of the year that was set to zero. The other year effects are deviations/differences from the year that was set to zero, The first option is only useful if you have only one fixed effect (the general mean will be in the year effects). The third option is relatively the most complicated, but it can be handy to have all year effects sum to zero.

Working out the third option in more detail gives:

We want to find $\hat{b} = (X'X)^{-1}X'Y$ - first $(X'X)^{-1}$ then $X'Y$: ($^{-1}$ refers to "inverse"

$$X' \quad X \quad = \quad X'X$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 5 \end{pmatrix}$$

$$X' \quad Y \quad = \quad X'Y$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 354 \\ 251 \\ 327 \\ 328 \\ 301 \\ 270 \\ 330 \end{pmatrix} = \begin{pmatrix} 2161 \\ 275 \\ 896 \end{pmatrix}$$

$$\hat{b} \quad = \quad (X'X)^{-1} \quad X'Y \quad = \quad \text{result}$$

$$\begin{pmatrix} \hat{b}_{mean} \\ \hat{b}_{1990} \\ \hat{b}_{1991} \end{pmatrix} = \begin{pmatrix} 0.1944 & -.0278 & -.1111 \\ -.0278 & 0.3611 & -.0556 \\ -.1111 & -.0556 & 0.2778 \end{pmatrix} \begin{pmatrix} 2161 \\ 275 \\ 896 \end{pmatrix} = \begin{pmatrix} 313 \\ -10.5 \\ -6.5 \end{pmatrix}$$

Summarizing the different options for X, and the resulting solutions:

General mean zero		First year zero (b1990=0)		Last year zero (b1992=0)		Sum of years to zero (b1990+ b1991 + b1992=0)	
X	$\hat{b}$	X	$\hat{b}$	X	$\hat{b}$	X	$\hat{b}$
1 0 0	302.5	1 0 0	302.5	1 1 0	330	1 1 0	313
1 0 0	306.5	1 0 0	4.0	1 1 0	-27.5	1 1 0	-10.5
0 1 0	330	1 1 0	+27.5	1 0 1	-23.5	1 0 1	-6.5
0 1 0		1 1 0		1 0 1		1 0 1	
0 1 0		1 1 0		1 0 1		1 0 1	
0 1 0		1 1 0		1 0 1		1 0 1	
0 0 1		1 0 1		1 0 0		1 -1-1	
μ = 0		μ = 302.5		μ = 330		μ = 313	
1990 = 302.5		1990 = 0		1990 = -27.5		1990 = -10.5	
1991 = 306.5		1991 = +4		1991 = -23.5		1991 = -6.5	
1992 = 330		1992 = +27.5		1992 = 0		1992 = 17	

We see from the different restrictions that the important parameters (the actual year differences) are always the same. In fact, with only one fixed effect in the model, these year differences can be estimated from the raw means for each year.

The story is different if we have more than one fixed effect. Suppose we now consider also the sex effect on yearling weight. We want now an estimate for the year effects, but also for the sex effect. Estimates of one fixed effect should be corrected for the other fixed effect. If in a particular year there are more males than females, we should account for that if estimating the year effects. Such non-balanced cases require really the power given by matrices and linear models.

With two fixed effects, we have to use two restrictions to obtain estimates. We will use the restriction that the solution of females in year 1992 is equal to zero (i.e. they represent the general mean).

The X matrix, and the solution become:

X	$\hat{b}$	meaning
$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 285.7 \\ -5.3 \\ -1.3 \\ 44.3 \end{pmatrix}$	the mean of females in 1992 the effect of year 1990 (relative to 1992) the effect of year 1991 (relative to 1992) the effect of males (relative to females)

Notice that the effect of year 1992 is greatly reduced because now we know that there was only an observation on a male. The difference between males and females was estimated with information from the previous years. In the first analysis we thought that 1992 was a particularly good year, but after consideration of the sex effect we know that the mean was only higher because there were relatively more males than females in 1992. Notice also that the difference between 1990 and 1991 has not

changed. This is because within these years there were equal numbers of males and females. This is indicated as a balanced design. Correcting for other fixed effects has therefore only an effect if those other effects are unequally contributing to a fixed effect under consideration. Only in a balanced design are the estimates of the different levels of a certain effect equal to the differences between the raw means of these levels. In practice we hardly ever have a balanced design, and we need a linear model to correct appropriately for all other effects.

The same example in ASREML:

Datafile: exmp2.dat

```
1990  Male    354
1990  Female  251
1991  Male    327
1991  Female  328
1991  Male    301
1991  Female  270
1992  Male    330
```

ASREML file: exmp2.as

```
analysis of test data 2 LM course
  year 3 !A
  sex 2 !A
  weight
exmp2.dat
weight ~ mu sex year
```

Output: exmp2.sln

```
  year                1990                  0.000       0.000
  year                1991                  4.000       33.92
  year                1992                  5.333       50.56
  sex                 Male                  0.000       0.000
  sex                 Female               -44.33       31.98
  mu                             1           324.7       31.98
```

And with forcing the sum of year solution to zero:

ASREML file: exmp2.as

```
analysis of test data 2 LM course
  year 3 !A
  sex 2 !A
  weight
exmp2.dat
weight ~ mu con(sex) con(year)
```

Output: exmp2.sln

```
  con(year)           1990                 -3.111       24.13
  con(year)           1991                 0.8889       21.32
  con(sex)            Male                  22.17       15.99
  mu                             1           305.6       18.07
```

Connectedness

A lack of connectedness among subclasses of fixed effects in a model can have serious consequences on estimability. If all subclasses of the fixed effects are full, i.e. contain at least one observation, then the data are completely connected and there are no problems with estimability. However, when several subclasses are empty the subclasses are not connected and some functions of b may not be estimable.

Connectedness can be evaluated by making tables of one fixed effect vs. another fixed effect and write the number of observations. For example:

Year \ sex	Male	Female
1990	1	1
1991	2	2
1992	1	0

Although not all subclasses are filled, the data is connected. It would not if the Male in the 1992 would be castrated such that we would have 3 sex classes, as in that case there would be a disconnected subset.

Year \ sex	Male	Steer	Female
1990	1	0	1
1991	2	0	2
1992	0	1	0

If there is disconnectedness in the data, the statistical programs will generally simply give no, or a zero solution to the effect associated with the disconnected subclass (i.e. no solution for year 1992 and Steer). Sometimes certain effects are *nested* within other effects. For example, herd 1 has only date from 1990 and 1991 whereas herd 2 has only data from 1992 and 1993. In that case the herd effect can not be estimated when years are fitted. When undertaking data analysis, it is important to understand such aspects of the design. For example, one could find out (e.g. with *awk)* how many year effects are in the data as well as how many year*herd combinations there are. If this is equal we know that one effect must be nested within the other.

Confounding

The best design to estimate parameters is a balanced design. There is an estimation problem if the data is disconnected. For example, in the last Table we can not distinguish between the effect of year 1992 and the effect of steers. However, in many cases the data is not balanced, but also not disconnected. Hence, there is a certain degree of confounding. Look at the following examples 3 and 4 and decide whether or not the fixed effects are significant.

Exmp3.dat

```
1990  Male    316
1990  Female  314
1990  Male    312
1990  Male    324
1991  Female  311
1991  Male    312
1991  Female  293
1991  Female  304
```

```
model statement: weight ~ mu con(sex) con(year)
```

Output: exmp3.asr

```
6 con(year)    1        2.06        2.06   5.806     [DF F_i F_a SED]
5 con(sex)     1        4.36        1.19   5.806     [DF F_i F_a SED]
```

```
model statement: weight ~ mu con(year) con(sex)
```

Output: exmp3.asr

```
6 con(sex)     1        1.19        1.19   5.806     [DF F_i F_a SED]
5 con(year)    1        5.23        2.06   5.806     [DF F_i F_a SED]
```

Exmp4.dat

```
15  109       287
17  116       298
18  119       306
18  116       303
19  117       302
19  119       312
20  121       316
21  122       324
```

```
analysis of test data 4 LM course
  age
  height
  weight
exmp4.dat
weight ~ mu height age
```

Output: exmp4.asr

```
  1 age        1        3.03        3.03 [DF F_inc F_all]
  2 height     1       70.50        1.67 [DF F_inc F_all]
```

```
analysis of test data 4 LM course
  age
  height
  weight
exmp4.dat
weight ~ mu age height
```

Output: exmp4.asr

```
  2 height      1        1.67        1.67 [DF F_inc F_all]
  1 age         1       71.87        3.03 [DF F_inc F_all]
```

The conclusion is that an inappropriate design does not allow you to make clear inferences about the different fixed effects. This might be ok if fixed effects are just 'nuisance parameters, e.g. when you are mainly interested in genetic parameters of EBVs, and fixed effects need to be corrected for. However, even in those cases, inadequate designs make estimates of fixed effects not very accurate. In example 3, the sex difference is estimated based on one comparison in each year (what is the female in 1990 happened to be a good one?) Inaccurate fixed effect estimates do affect accuracy of genetic parameter estimates as well.

Hypothesis testing

Requirements: Assume that y has a multivariate distribution. Hypothesis testing requires knowing the distributions of sums of squares. A sum of squares, say y'Qy, will have a chi-squared distribution if QV is idempotent (i.e. this matrix times itself is equal to itself) and if y is MVN.

The most relevant sums of squares are:

$SS_{Total} = y'Q_Ty$ where $Q_T = V^{-1}$

and $SS_{model} = y'Q_Ry$ where $Q_R = V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}$

and $SS_{Residual} = y'Q_E$ where $Q_E = V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}$

It can be proven that SSR and SSE are independent ch-square variables ($Q_RVQ_E = 0$).

Testing the model:

The ratio of two independent central chi-square variables has an F-distribution. The adequacy of the whole model is tested as:

$$F_M = \frac{SSR/r(X)}{SSE/(N-r(X)}$$

where r(X) is the rank of X and N is the total number of observations.
The whole model is usually significant as it contains the mean (which is usually significantly different from zero). It is more useful to test subsets of the parameter vector b. Various functions of b can be tested. An hypothesis test consists of

1. The null hypothesis
2. the alternative hypothesis
3. a test statistic
4. a probability level or rejection region

The null hypothesis can be written as

H'b = c

Or: H'b-c = 0

Where:

H must be of full rank

H'b must be an estimable function

If these conditions are met, H'b is testable. The test statistic is

$$F = \frac{s/r(H')}{SSE/(N-r(X))}$$

where $s = (H'\beta\text{-}c)'(H'CH)^{-1}(H'\beta\text{-}c)$

and $C = (X'V^{-1}X)^{-1}$

Example:

Example data set 5

Calf ID	**Age of Dam (yr)**	**Breed**	**Growth Rate (kg/day)**
1	2	AN	2.10
2	3	AN	2.15
3	4	AN	2.20
4	5+	HE	2.35
5	5+	HE	2.33
6	2	HE	2.22
7	3	HE	2.25
8	3	HE	2.27
9	4	SM	2.50
10	5+	SM	2.60
11	2	SM	2.40
12	2	SM	2.45

An appropriate model to describe this data would be a two-way cross classified model without interaction:

$$y_{ijk} = b_0 + b_i + b_j + e_{ijk}$$

where

y_{ijk} is an observation on the growth rate of calves

b_0 is the overall mean

b_i is an effect due to the age of dam of the calf (i = 1,…4)

b_j is an effect due to the breed of the calf (j=1,….3)

e_{ijk} is the residual for each observation

The model written in matrix notation is

$$y = Xb + e$$

The assumptions of the model are

- there are no breed by age of dam interactions
- all other effects were the same for all calves, eg. diet, age, cg
- errors terms are independent and random variables identically distributed around a mean of 0 and a variance of σ_e^2.

The expectation of y is

$$E(y) = Xb$$

and the variance of y is

$$V(y) = I\,\sigma_e^2$$

where

$$Xb = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \\ \beta_{21} \\ \beta_{22} \\ \beta_{23} \end{bmatrix}$$

Normal Equations

The normal equations for GLS are

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

however as $V(y) = I\,\sigma_e^2$ then

$$\sigma_e^{-2}\,(X'X)b = \sigma_e^{-2}\,X'y$$

and the GLS equations reduce to those of OLS equations, ie.

$$(X'X)b = X'y$$

which in expanded matrix form is

$$\begin{bmatrix} 12 & 4 & 3 & 2 & 3 & 3 & 5 & 4 \\ 4 & 4 & 0 & 0 & 0 & 1 & 1 & 2 \\ 3 & 0 & 3 & 0 & 0 & 1 & 2 & 0 \\ 2 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 3 & 0 & 0 & 0 & 3 & 0 & 2 & 1 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 5 & 1 & 2 & 0 & 2 & 0 & 5 & 0 \\ 4 & 2 & 0 & 1 & 1 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \\ b_{21} \\ b_{22} \\ b_{23} \end{bmatrix} = \begin{bmatrix} 27.82 \\ 9.17 \\ 6.67 \\ 4.70 \\ 7.28 \\ 6.45 \\ 11.42 \\ 9.95 \end{bmatrix}$$

Obtaining Solutions

X'X is a positive semi-definite matrix with a rank of 6. The dependencies are that columns 2, 3, 4 and 5 and then columns 6, 7 and 8 both sum to give column 1 and thus two constraints on the solution are needed. Letting b_0 and b_{11} be then set to zero, a generalised inverse of X'X is equal to $(X'X)^-$

$$\begin{bmatrix} 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.714 & 0.225 & 0.322 & -0.313 & -0.414 & -0.137 \\ 0.000 & 0.000 & 0.225 & 0.793 & 0.194 & -0.339 & -0.167 & -0.247 \\ 0.000 & 0.000 & 0.322 & 0.194 & 0.670 & -0.172 & -0.396 & -0.216 \\ 0.000 & 0.000 & -0.313 & -0.339 & -0.172 & 0.551 & 0.194 & 0.128 \\ 0.000 & 0.000 & -0.414 & -0.167 & -0.396 & 0.194 & 0.524 & 0.141 \\ 0.000 & 0.000 & -0.137 & -0.247 & -0.216 & 0.128 & 0.141 & 0.366 \end{bmatrix}$$

for which the corresponding solution vector is $b = (X'X)^-X'y$;

$$\begin{bmatrix} b_0 \\ b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \\ b_{21} \\ b_{22} \\ b_{23} \end{bmatrix} = \begin{bmatrix} 0.000 \\ 0.000 \\ 0.052 \\ 0.082 \\ 0.147 \\ 2.105 \\ 2.204 \\ 2.430 \end{bmatrix}$$

However G above is one of several generalised inverses for X'X and thus there are several possible solution vectors. In fact there are an infinite number of possible solution vectors which are given by the formula,

$$b^o = (X'X)^-X'y + (I - (X'X)^-X'X)z$$

where *z* is an arbitrary vector of constants.

Properties of Solutions

While there are an infinite number of different solution vectors to the GLS equations the sum of squares due to the model is unique.

$$\begin{aligned} SSR &= \mathbf{b'X'y} \\ &= 64.737 \\ &= \mathbf{b^{o}'X'y} \\ &= y'X'(X'X)^-X'y \\ &= 64.737 \end{aligned}$$

Therefore when it comes to testing the model, the values for mean squares and F-tests are likewise independent of the solution vector.

Estimable Functions

By computing the expected value of the solution vector, the *functions* of true parameters that have been estimated by a particular generalised inverse can be determined. These solutions are estimable because *the solution vector is a linear function of y which is always estimable.*

Estimable functions are unique regardless of the solution vector. Consider the function b_{12} - b_{11} (this function is obtained by multiplying the third row of the matrix of estimable function by b), which can be more generally written as

$$k'b = (0\ \ -1\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0)b = b_{12} - b_{11}$$

If another solution vector is used, the same value will be produced for the same function. Thus one quick way to determine if a function is estimable is to multiply it by b and b^o; if the results differ then that function is not estimable. A further method to determine if a function is estimable is to check if

$$k_i'(X'X)^-X'X = k_j'$$

Variance of Estimable Functions

The variance of an estimable function is given as

$$\begin{aligned} V(k'b) &= k'V(b)k \\ &= k'V((X'X)^-X'y)k \\ &= k'(X'X)^-X'V(y)XG'k \\ &= k'(X'X)^-X'X(X'X)^{-'}k\ \sigma_e^2 \end{aligned}$$

and since $k'(X'X)^-X'X = k'$, if k is estimable

$$= k'(X'X)^-k\ \sigma_e^2$$

So when k' = (0 -1 1 0 0 0 0 0) and k'b = 0.052 then V(k'b) =

$$\begin{pmatrix}0 & -1 & 1 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\begin{bmatrix}0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000\\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000\\ 0.000 & 0.000 & 0.714 & 0.225 & 0.322 & -0.313 & -0.414 & -0.137\\ 0.000 & 0.000 & 0.225 & 0.793 & 0.194 & -0.339 & -0.167 & -0.247\\ 0.000 & 0.000 & 0.322 & 0.194 & 0.670 & -0.172 & -0.396 & -0.216\\ 0.000 & 0.000 & -0.313 & -0.339 & -0.172 & 0.551 & 0.194 & 0.128\\ 0.000 & 0.000 & -0.414 & -0.167 & -0.396 & 0.194 & 0.524 & 0.141\\ 0.000 & 0.000 & -0.137 & -0.247 & -0.216 & 0.128 & 0.141 & 0.366\end{bmatrix}\begin{bmatrix}0\\-1\\1\\0\\0\\0\\0\\0\end{bmatrix}\sigma_e^2$$

$$= 0.714\ \sigma_e^2$$

Similarly if a number of estimable functions are to be considered then

$$K' = \begin{bmatrix}0 & -1 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0\\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0\end{bmatrix}$$

then

$$K'b = \begin{bmatrix}0.082\\0.147\\2.105\end{bmatrix}$$

and

$$V(K'b) = K'(X'X)^{-}K\,\sigma_e^2 = \begin{bmatrix}0.793 & 0.194 & -0.339\\ 0.194 & 0.670 & -0.172\\ -0.339 & -0.172 & 0.552\end{bmatrix}\sigma_e^2$$

The standard errors of the estimable functions are obtained as the square root of the variances of the estimable functions located on the diagonals above.

In ASREML you can use '*contrast*' to test hypothesis.

Least Square Means

Least square means (LSM) are commonly used in scientific articles as they relate directly to the actual measurements of data and are thus readily understood. However least square means are not equal to the actual raw means but are estimable functions and as such are, of course, unique. In fact LSMs are simply estimators of the marginal

means of different classes or subclasses that would be expected in a balanced design, for example

	sex 1	**sex 2**	**LSM (**Year**)**
Year 1	11.0	9.0	10.0
Year 2	16.0	12.0	14.0
LSM (sex**)**	13.5	10.5	12.0

Here the LSM for sex 1 corrected for year effects is 13.5 or alternatively the LSM for Year 1 is 10.0 corrected for sex effects. In the same way least square means can be derived for the different sub-classes of each effect in the previous data corrected for all other effects through the use of regression coefficients. For instance the least square means for the different levels of the effect of age of dam would be

Age of Dam	**L.S. Mean**
2	2.247
3	2.299
4	2.329
5+	2.394

These are not simply $\beta_0 + \beta_{1i}$, especially in this case as this is not an estimable function. Instead these means are estimating $\beta_0 + \beta_{1i} + 1/3(\beta_{21} + \beta_{22} + \beta_{23})$. Alternatively the least square means for the different levels of breed effects are

Breed	**L.S. Mean**
Angus	2.176
Hereford	2.275
Simmental	2.501

which are estimating $\beta_0 + 1/4(\beta_{11} + \beta_{12} + \beta_{13} + \beta_{14}) + \beta_{2i}$.

Analysis of Variance

For the example data, the basic analysis of variance table is

Source	d.f	Sums of Squares	Means Square	F - stat
Total (SST)	N-1 = 12	y'y = 64.740	5.3950	10295.8
Mean (SSM)	1	$N\bar{y}^2$ = 64.496	64.4960	123084.0**
Model (SSR)	r = 6	b'X'y = 64.7371	10.7895	20590.6**
Residual (SSE)	N - r = 6	y'y - b'X'y = 0.0031	0.0005	

The estimate of the error variance is given by the residual means square (MSE). The test for the adequacy of the model is given by the calculation of an F-statistic for the model:

$$F_M = MSR/ MSE$$

which at $P < 0.05$, is highly significant for this example. A significant F_M indicates that the solution vector is not a null vector and that the model explains some of the major sources of variation. This is generally significant because the solution vector includes the mean of the observations, which is usually different from zero. Alternatively the multiple correlation coefficient, R^2 can be used to determine the amount of variation accounted for by the model where

$$\begin{aligned} R^2 &= (SSR - SSM)/(SST - SSM) \\ &= 0.987 \end{aligned}$$

The higher the value for R^2 the better the model, which in this case is a very adequate description of the variation in calf growth rates.

Hypothesis Tests

As mentioned above the test of a models appropriateness is usually significant because the solution vector for the model includes the mean of observations. Therefore when testing the importance of a model, it is of greater worth to test the significance of the elements of *b* other than the mean.

The General Linear Hypothesis Procedure

Using the generalised linear hypothesis procedure, the test of SSR is directly partitioned into sub-hypotheses which test the various estimable functions of *b*. For the current example the two tests of interest would be

i) age of dam effects
ii) breed of calf effects

For the general linear hypothesis, the null hypothesis for testing age of dam effects would be $H_1'b = 0$ where

$$H_1' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$

and

$$H_1'b = \begin{bmatrix} b_{11} - b_{12} \\ b_{11} - b_{13} \\ b_{11} - b_{14} \end{bmatrix}$$

For H'b to be testable

i) H' must have full row rank
ii) H'b must be an estimable function (determined using methods described previously)

and since these conditions hold, $H_1'b$ is testable.

The sum of squares for $H_1'b$ is

$$s_1 = (H_1'b)'(H_1'(X'X)^- H_1)^{-1} H_1'b$$

and given

$$H_1'b = \begin{bmatrix} -0.052 \\ -0.082 \\ -0.147 \end{bmatrix}$$

then

$$s_1 = 0.0357$$

with 3 degrees of freedom, ie. n_1 - 1. The F-test is

$$F_1 = \frac{s_1 / r(H_1')}{\sigma_e^2}$$

$$= 22.7$$

which at the 0.05 level means that the differences among age of dam groups is significantly different from zero and accounts for some of the variation explained by the model.

Similarly to test breed of calf differences

$$H_2' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix},$$

$$H_2'b = \begin{bmatrix} -0.099 \\ -0.325 \end{bmatrix}$$

and the sums of squares for breed effects = 0.174 with 2 degrees of freedom. In summary of the partitions of SSR for calf growth data

Source	**d.f.**	**Sum of Squares**	**Means Square**	**F-test**
Model	6	64.737	10.7895	20590.6
Mean	1	64.496	64.4960	123084.0
Age of Dam	3	0.0357	0.0119	22.7**
Breed of Calf	2	0.174	0.0870	166.0**

which shows that both elements of the model are significantly different from zero and both explain some of the variation in calf growth rates.

Reduction Notation

Another means of testing the appropriateness of a model and its elements is to examine the significance of the reduction in sums of squares after regressing y on each element of the model separately. In the current example, the full model could be written in component form as

$$y = \mu 1 + X_1 b_i + X_2 b_j + e$$

with the different sub-models being

$$\text{Model 1: } y = \mu 1 + X_2 b_j + e$$
$$\text{Model 2: } y = \mu 1 + X_1 b_i + e$$
$$\text{Model 3: } y = \mu 1 + e$$

The notation for the reductions of these models are:

$R(\mu , b_i , b_j)$ = the sums of squares due to fitting the full model
$R(\mu , b_j)$ = the reduction sums of squares due to fitting model 1
$R(\mu , b_i)$ = the reduction sums of squares due to fitting model 2
$R(\mu)$ = the reduction sums of squares due to fitting the mean

and for each of these models reductions in sums of squares for each of these models would be obtained by constructing OLS equations and then solving to give

$R(\mu , b_i , b_j) = 64.737$
$R(\mu , b_j) = 64.701$
$R(\mu , b_i) = 64.563$
$R(\mu) = 64.496$

To test the null hypothesis $b_j = 0$, ie. that the differences in age of dam effects are not significantly different from zero

$$\begin{aligned} s_1 &= R(\mu, b_i, b_j) - R(\mu, b_j) \\ &= R(b_i \mid \mu, b_j) \\ &= 0.0357 \end{aligned}$$

with 3 degrees of freedom. Similarly for the differences in breed of calf

$$\begin{aligned} s_2 &= R(\mu, b_i, b_j) - R(\mu, b_i) \\ &= R(b_j \mid \mu, b_i) \\ &= 0.174 \end{aligned}$$

with 2 degrees of freedom. Finally to test the significance of the model after correcting for the mean gives

$$\begin{aligned} s_3 &= R(\mu, b_i, b_j) - R(\mu) \\ &= R(b_i, b_j \mid \mu) \\ &= 0.241 \end{aligned}$$

While these results are the same as those obtained from the general linear hypothesis procedure, the method requires that individual analyses of variance be performed for each effect. In this way, the former technique is a much simpler and easier method in that the sums of squares for fixed effects and their significance in the model can be determined with greater efficiency.

SUMMARY

The general aim of estimation is to obtain the highest quality estimates of parameters for a particular model. When the assumptions of a model are valid the estimates can be considered as *best, unbiased* and if a linear model is fitted then *linear,* ie. BLUE.

OLS - the model assumes that error terms are independent and they have a common variance, ie. $V = I_N \sigma_e^2$ which is a diagonal matrix for which all non-zero terms are equal.

WLS - equivalent to OLS except that the values along the diagonal in V may differ due to, for example, the number of records.

GLS - equivalent to WLS except that V is a VCV matrix of error terms in which diagonal terms will differ due to differing error variances for each observation and *ij*th elements which are the covariances between observations

Note that WLS and OLS are simply special cases of GLS.

Exercises for linear models

1) An introductory example

To get some feel for why it is useful to calculate sums of square in the construction and testing of statistical models for prediction, consider the following example.
Suppose we have 4 observations and a one-way classification with 2 levels (A and B).
Calculate the sum of squares for the total, the mean, the model and the residual.
Residual 1 refers to a model where only the mean is fitted and residual 2 to a model where also the class effect is fitted.
Calculate sums of squares 'by hand' based on the numbers the column.

	class	Observation	Mean	Residual 1	Predicted Y	Residual 2
	A	8				
	A	9				
	B	11				
	B	12				
Sum of Squares						

2) Regression Models:

We have measured the litter size of a group of sows, and are interested in some effects on this trait, in particular the effect of the age of the sow, and the effect of fat depth at insemination.

<u>1 single regression:</u>

y = litter size pigs [7 8 9 8 9 10 9 10 11 12]
x = sow weight at insemination (kg) [100 110 120 125 125 130 130 145 150 160]

<u>2 multiple regression</u>

y = as before, x1 = as x before
x2 = fat depth at insemination (mm) [20 30 25 40 25 30 35 40 35 35]

Estimate regression coefficients for linear regression models

Test whether regression coefficients are significantly different from zero.

(You may try also 2nd order regression if you like)

Compare estimates of single and multiple regression.

3) **Linear models with class variables**

1 One way classification

Given are data on three pig breeds. Estimate the breed effect

Yorkshire	Landrace	Pietrain
800	600	600
700	700	
600		

Test whether the breed effect is significant

Test whether the Yorkshire breed is significantly different from Landrace.

2 Two way classification

Given are data of daily growth of cattle in an experiment, where 3 feeding levels were tested. Given are observations on some bulls.

		Feeding level		
		Pasture Queensland	Feedlot	Pasture Victoria
Breed	Angus	200, 165	450, 460	300, 350
	Hereford	220	426, 390, 430	310, 320, 330
	Brahman	260, 240, 235	380, 450	280
	St. Gertrude	245, 220, 250,240	420, 440,	300

In a one-way analysis:

Analyze the effect of breed. Test the general effect

Analyze the effect of feeding regime. Test general effect.

Repeat the previous analysis in a 2-way classification.

Regression and class effects

Consider the following data where fat depth was measured on bulls in two feeding regimes. The bulls were measured at different ages.

Fat Depth (mm)	Feeding Regime	Age at measureing (Mo)
20	Intensive	10
20	Int	14
19	Int	15
24	Int	16
24	Int	17
25	Int	20
26	Int	20
19	Extensive	17
19	Ext	19
21	Ext	21
20	Ext	23

1) Estimate (and test) the effect of age on Fat Depth without consideration of feeding regime
2) Estimate the same effect with consideration of feeding regime

Further revision questions

1. Define:

   Fixed effects

   Random effects

2. Write a fixed effect model for two independent variables

   Give the expectation of the dependant variable (1^{st} moment)

   Give the variance of the dependant variable (2^{nd} moment)

   State the assumption of the model

3. What are the differences between:

   ordinary least squares estimates

   weighted least squares estimates

   generalised least square estimates

Mixed Models

The majority of the previous work in this session has examined the estimation of fixed effects using linear regression models. However in animal breeding the prediction of random effects for individual animals and their variance for a population is of more value.

As presented previously the mixed linear model in matrix form is

$$y = Xb + Zu + e$$

Recall that G is the VCV matrix of *u* and R is the VCV matrix of *e* such that

$$V = V(y) = ZGZ' + R \qquad ...(8)$$

Note that if R was reduced to its simplest form, namely $I\sigma^2_e$ and *u* was ignored, the mixed model equation would reduce to the standard linear model (5).

If G and R are known, estimates of *b* and the predicted value of *u* are

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

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

which as a result of V given in (8) means that these effects have been estimated simultaneously and thus

i) β is the GLS solution for *b* as well as its best linear unbiased estimator (BLUE)

ii) $\hat{u}$ is the best linear unbiased predictor (BLUP) of *u*

iii)

Henderson(1959) developed a set of equation that simultaneously generate BLUE(Xβ) and BLUP(u), these equation being called mixed model equations: MME.

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

Hence , there are two subsets of equations:

$$X'R^{-1}X\beta + X' R^{-1} Z \hat{u} = X' R^{-1}y$$
$$Z'R^{-1}X\beta + (Z' R^{-1} Z+G^{-1})\hat{u} = Z' R^{-1}y$$

Substituting for $\hat{u}$ gives

$$X'R^{-1}X\beta+X' R^{-1}Z(Z' R^{-1}Z+G^{-1})^{-1}Z' R^{-1}(y-X\beta)=X' R^{-1}y$$

To be short: $X' V^{-1}X\beta = X' V^{-1}y$

where $V^{-1}= R^{-1}- R^{-1}Z(Z' R^{-1}Z+G^{-1})^{-1}Z' R^{-1}$

Checked by VV^{-1}

Hence, in the MMM we estimate

$BLUE(b) = \beta = (X' V^{-1}X)^{-1} X' V^{-1}y$ is a GLS estimate

$BLUP(u)= \hat{u} =(Z' R^{-1}Z+G^{-1})^{-1}Z' R^{-1}(y-X\beta)$

It is interesting to note however that with $V = I\sigma^2$ the solutions to these mixed model equations are merely least square estimates.

Variance of predictors and prediction errors

A prediction from a mixed model uses a combination of estimates of fixed effects and predictions of random effects. For example, we can predict the performance of a certain daughter of a bull in a certain herd at a certain age.

The *predictand* is K'b + M'u
The *predictor* is a linear function of y, i.e. L'y (practically, a linear combination of the estimated parameters, which in themselves are linear functions of the data)

The Prediction error is the difference between the predictor and the predictand, i.e. K'b + M'u – L'y. If this is zero, then the prediction is unbiased.

The prediction error variance:

$V(b - \beta) = V(\beta) = (X'V^{-1}X)^{-}$
$V(u - \hat{u}) = V(\hat{u}) + V(u) - 2Cov(u, \hat{u})$

$= V(u) - V(\hat{u})$ as $Cov(u, \hat{u}) = V(\hat{u})$

$= G - V(\hat{u})$

Further: $Cov(\beta , u - \hat{u}) = 0$, and $Cov(\beta , \hat{u}) = 0$

These PEVs can best be obtained from the mixed model equations. The solutions to the MME can be written as

$$\begin{bmatrix} \beta \\ \hat{u} \end{bmatrix} = \begin{bmatrix} C_{XX} & C_{XZ} \\ C_{ZX} & C_{ZZ} \end{bmatrix} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Where the matrix is the generalized inverse of the coefficient matrix of the MME, i.e. C_{XX} is the 'fixed effects part' of the inverse, and NOT the inverse of $XR^{-1}X$

Now, $Var(\beta) = C_{XX}$

$V(\hat{u}) = G - C_{ZZ}$

And $V(u - \hat{u}) = C_{ZZ}$

PEV's of estimated breeding values

In a BLUP model we can have animals' additive genetic effects as random effects. Now the $V(\hat{u})$ = the variance of the EBV's. From quantitative genetic theory we know that $var(EBV) = r_{IH}^2 V_A$, where r_{IH} is the accuracy of the EBV and V_A is the additive genetic variance. From the BLUP model we can first obtain the diagonal element of the inverse C_{ZZ} (sometimes we approximate this value as the MME coefficient matrix is often not inverted), for animal I this is C^{ii}.

The Prediction Error Variance of the EBV: $PEV = C^{ii}$

This is also equal to $(1 - r_{IH}^2) V_A$

Note that var(EBV) + PEV add to VA

Note that MME often have multiplied out R^{-1}, i.e. we use Z'Z rather than $Z'R^{-1}Z$ etc. In that case

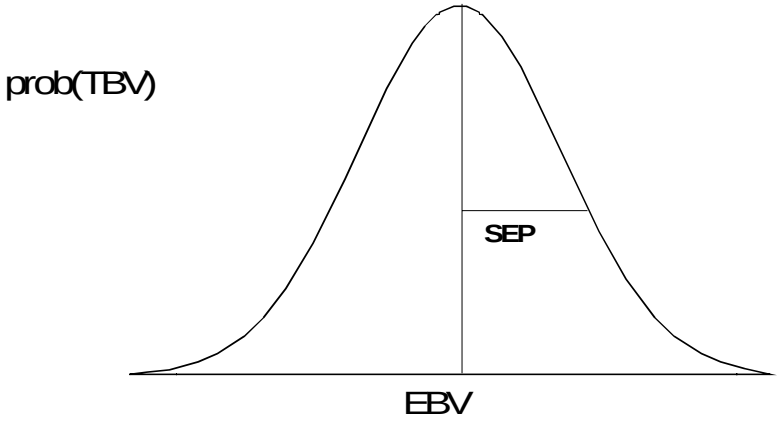
The Prediction Error Variance of the EBV: $PEV = C^{ii}\sigma_e^2$

And the Standard Error of Prediction (SEP) is $\sqrt{}$(PEV). ASREML gives SEP values behind the solution of random effects (*.sln file)

Think again about the extreme cases:

- when there is no information, and accuracy is 0: all EBV's will then be 0 and the variance of the prediction error is equal to V_A.
- when there is full information, the EBV will be equal to the true BV and the variance of the prediction error will be 0.

The prediction error of an EBV is important since it gives us a clue of how far the true breeding



value could be off the EBV. This is important for example to answer questions like: how much could an EBV still change if we obtain more information on the animal. Changes in EBV's are not good for the industry's confidence in the genetic evaluation system. However, we have to realise that an EBV is never exact, unless the accuracy is 100%. We expect the true breeding value to be the same as the EBV, but there is a certain probability that it will be a bit different. The probability distribution of the true EBV, given an EBV looks like in the figure.

Hypothesis Testing in Mixed Models

Hypothesis testing in the case of mixed models with unbalanced data is not well understood. Many analyse the fixed effects only and ignore random effects. Other treat random effects (e.g. sires) as fixed. In hypothesis testing, expectations are derived assuming the true model. However, the variance components needed in a mixed model are estimates, and therefore strictly solutions for fixed effects (combinations) are not BLUE.

If G and R are known,
then V is also known and estimates of b are BLUE and the hypothesis test is as described before. To test H'b-c = 0

We used the test statistic $F = \dfrac{s / r(H')}{SSE/(N - r(X))}$

where $s = (H'\beta\text{-}c)'(H'CH)^{-1}(H'\beta\text{-}c)$ and $C = (X'V^{-1}X)^{-1}$

This test is exact and best, given that G and R and known, or known to proportion.

When G and R are not known,
there is no best test and BLUE of b is not possible. If estimates of G and R are used, then hypothesis testing is only approximate. The possibilities are:

a) Estimation by computing as though random effects were fixed:

$$\begin{bmatrix} \beta \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}^{-} \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

For this case, if K'b is estimable, K'β is an *unbiased* estimator of k'b., however, it is not a minimum variance estimator of b and it does not have maximum power.
An exact test is given by: $Q/f\,\hat{\sigma}_e^2 \sim F[f, N\text{-}r(X,Z)]$
for $Q = (K'\beta - m)'(K'C_{11}K)^{-1}(K'\beta m)$ and $\hat{\sigma}_e^2$ is the estimated residual variance from the model.

If K'b is not estimable, no exact test exists. The estimate k'b depends on the choice of m'u, and the F-test can be inflated because the denominator (containing only residual variance) is too small. The degree of bias in F depends on the ratio of variance components.

b) Estimation ignoring all random effects

$\beta = (X'X)^{-}X'y$

If K'b is estimable, K'β is an *unbiased* estimator of k'b., however, it is not a minimum variance estimator of b and it does not have maximum power. No exact test is possible. F-tests are often inflated if K'b is not estimable. An approximate test can improve the properties of the test but few statistical packages would accommodate this.

c) Estimation by computing with estimates of the variances of the random effects

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

This approach gives unbiased estimates and often with smaller sampling error than when treating random effects as fixed. The F test is approximate, with better properties if the estimated value for G approximates the true value. Effectively the residual variance is corrected for random effects. The denominator of the F-test contains residual variance and a term for the variance components. This F-test is more precise and the denominator is not inflated as in a) or b). If no good estimates of variance components exist, however, it might be safer to follow approach 1)

d) Simultaneous estimation of fixed effects and variance components for random effects, e.g. using REML. This gives an approximated F-test as in c). This is often the most sensible approach, given that there is a reasonable amount of data to estimate variance components. ASREML will provide the most appropriate F statistic for this case.

Example/Exercise:

Consider the following data set with two treatments measured on 4 cows (2 for each treatment) with 5 repeated measurements per cow (i.e. 10 measurements per treatment.

Treatment	Cow	
I	A	451; 456; 462; 449; 455;
	B	472; 469; 476; 467; 462;
II	C	481; 475; 482; 489; 483;
	D	510; 502; 499; 507; 501;

Test the treatment effect, with and without cow fitted.

RECOMMENDED (backup) READING

Kennedy, B. 'old'. Linear Models - Course notes. In *AGBU library*

Schaeffer, L.R. 1993. Linear Models and Computing Strategies in Animal Breeding - Course notes. In *AGBU library*

Quaas, R.L., Anderson, R.D. and Gilmour, A.R. 1984. *BLUP School Handbook - Use of mixed models for prediction and for estimation of (co)variance components.* AGBU, Armidale.

Mrode, R.A. 1996. *Linear Models for the Prediction of Animal Breeding Values.* CAB International, Wallingford.

Neter, J., Wasserman, W. and Kutner, M. 1985. *Applied Linear Statistical Models*. Irwin, Illinois.

Introduction to MATLAB

Matlab is a program that handles matrices. It is very convenient for small examples

MATLAB is case-sensitive, i.e. a variabele “x” is not the same as “X”

Statements can be made interactive, but it is also easy to make a program
Click on “File→New→Mfile”. This bring you in notepad. Save the file with extensuion M, and as “All Files” (NOT as a “Text File”)

A “%” (percent sign) means that you can make comments in that line, it is not executed
A “;” (semi-colon) at the end of the line means that MATLAB will not print the result
A matrix is given between [ ], and rows are separated by a semi column, or by a retrun
You can use submatrices within a matrix

Other ‘tricks’:

identity matrix of order n :	eye(n)	
Nr. of rows of a matrix	nr=size(A,1)	
Nr. of columns	nc=size(A,2)	
matrix (or vector) with ones only:	ones(n,m)	
inverse of matrix A	inv(A)	
multiplication	A*B	
transpose	A’	
only diagonals (or make diagonal)	diag(B)	
Add a column b to A	A=[A b]	
sub matrices	B=A(1:3,2:4)	matrix B has elements of rows 1 to 3 and rows 2 to 4 from matrix A.
	b=A(:,2)	b is equal to the 2nd column of A
	c=b([1,4,7])	b contains the 1st, 4th and 7th elem’s of b.
sums all elements of y	sum(y)	

» % program for a simple regression example

```
y=[74; 82; 84];
X=[1 160;
    1 170;
   1 180];
n=size(y);
```

```
% solutions:
b=inv(X'*X)*X'*y

% Analysis of Variance
% sums of squares for total
SST=y'*y

% sums of squares for mean
SSM=n*(sum(y)/2)^2                    %  x^2 takes the square of variable x

% sums of squares for model & regression
SSMod= y'*X*inv(X'*X)*X'*y;
SSA=SSMod-SSM

% sums of squares for residual
SSE=SST-SSMod
```