Chapter 3:
Introduction to Linear Models

Julius van der Werf

Linear models are commonly used to describe and analyse data in the biological and social 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

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. Problems 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 these comparable means. An estimate of a herd mean based on 5 animals is less accurate than one with 100 animals.

Hence, observed phenotypes need to be corrected for other non-genetic effects before they are used to estimate breeding value. These other effects need to be estimated in an unbiased way. Bias could occur, for example, if some herds use better sires than other herds. If we want to correct for the environmental effect of herds, we could not simply compare herd means, because unequal genetic means have to be taken into account. Also within herds or flocks effects of season, birth type, or age of measurement can influence the outcome of a measurement. A statistical procedure that allows unbiased estimation of a number of effects simultaneously is based on linear models. This simultaneous estimation is important for unbiased estimation of effects when different levels of one effect are not equally represented at all levels of another effect, i.e. when we have unbalanced date. Field data, often used for genetic evaluation and other quantitative genetic analyses, are rarely balanced. Linear models use matrices to layout the design in such data and, as we will show, prove to be very convenient in order to correct the different effects for each other.

Linear models form the basis of Best Linear Unbiased Prediction (BLUP), and are an important aspect of the method since it provides the machinery to correct breeding values for a number of systematic environmental effects (usually termed fixed effects) simultaneously. BLUP estimation of breeding values is based on a mixed model, which is a

linear model containing fixed effects as well as random effects (the additive genetic values), - see later in this course.

The purpose of this lecture is to familiarise you with linear models. Based on a simple example, we will present a linear model as the statistical method to estimate fixed effects. Understanding such examples is important for understanding how BLUP corrects for fixed effects in genetic evaluation procedures.

Simple Example

The main practical advantage of a linear model is that it can appropriately 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. Very sophisticated linear models can be formulated, accommodating different effects, and possibly their interactions, covariances between different effects, different types of distributions etc.

Linear Models in Genetic Evaluation

An important feature of genetic evaluation is that animals are compared fairly. When assessing the phenotypic performance of animals, unfair comparisons could be made if animals perform in different herds, are born in different years or seasons, have a different birth type (singles vs twins), are measured at different ages, etc. Before assessing the genetic merit of animals, such factors should be taken into account, i.e. we want to take an animal's performance after correcting it for such effects. A simple suggestion could be to take a performance as a deviation from a contemporary mean, i.e. a deviation from the mean performance of a group of animals that have performed under similar conditions, have the same sex, and the same age.

However, simply taking such deviations from the class means may give biased correction. Suppose that the sires used in herd A are on average superior to sires used in herd B. For an animal with a specific breeding value it is than more difficult to have a positive deviation in herd A than it is in herd B, even if the herds had exactly the same management and environmental conditions. Two animals with the same true breeding value would then have different EBV depending on the herd they were tested in. This is a drawback, since we don't want an EBV to be biased by such fixed effects.

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 in the late forties by C.R. Henderson (Henderson, 1973). The BLUP procedure takes account of such fixed effects, and is therefore a Best Linear Unbiased Prediction of the breeding value.

The ability to compare animals in different herds, and to correct for herd effects in an unbiased way, depends also on the structure of the data. To be able to compare animals across herds, herds will need to be linked.

The following example might illustrate the problem.

Suppose we have progeny means from 4 sires in 2 herds as follows:

Herd	Sire 1	Sire 2	Sire 3	Sire 4
1	325	275		
2			325	275

From within herd comparison we know that sires 1 and 3 are superior sires. But how do sires compare across herds? To be able to make such a comparison we will need a fifth sire that has been used in both herds. Such a sire would be a link sire or a reference sire. Suppose the link sire has progeny in the two herds as follows:

Herd	Sire 1	Sire 2	Sire 3	Sire 4	Sire 5
1	325	275			325
2			325	275	375

From this new information we know that the environmental effect of herd 2 must be better than of herd 1, as progeny from the same sire perform better in herd 2. Given that information, we can derive that Sire 1 must be a better sire than Sire 3, as he is able to have the same progeny mean, but in a herd that is not as good. Hence, to be able to compare all sires, all data from all herds and sires should be evaluated simultaneously.

In reality, different sires could also be mated to different cows, and the merit of cows can be worked out similarly, with the notion that cows are themselves daughters of sires that are used in different herds. It would be quite impossible to make fair comparisons among animals if not all data was analysed jointly, across herds, and a powerful method exists that can take into account how each observation is affected by a number of fixed effects.

In the next section, we will present a linear model and demonstrate that it is a powerful framework for unbiased estimation of fixed effects. It will show how different fixed effects can be corrected for each other in a straightforward approach. Using examples, we will indicate how to set up linear model equations. These principles are important for understanding BLUP.

Linear Models

Estimation of Fixed Effects

Linear models are used throughout to estimate different effects acting on observations. We introduce an example data set to illustrate the concept. The data consists of 7 weight observations on 7 animals, producing in 3 different years. Given are the animal IDs and their yearling weight (in brackets)

Year of Birth	Males	Females
2000	1 (354)	2 (251)
2001	3 (327) 5 (301)	4 (328) 6 (270)
2002	7 (330)	

Model with One Fixed Effect

Given certain data, we have to wok on model building. We are first interested in the effect of birth year on the weight of an animal. Estimation of year effects can be obtained from a statistical model where the fixed effect of Year in the example data set is fitted to explain variation in the data.

Weight = general mean + effect of year + random error

$y_{ij} = \mu + year_i + e_{ij}$

The y-variable is the dependent variable
The x-variable is the independent variable.
The unknown year effects ($year_i$) are model parameters.
The model proposed is a linear model as the expected value of y, E(y) is a linear combination of parameters.

In matrix notation:

$y = Xb + e$

where y is a vector with observations on the weight of an animal,
b is a vector with the different year effects,
X is an 'incidence matrix', indicating which observation was observed in which year, and
e is a vector with residual effects

The model becomes

$$y = X b + e$$

$$\begin{pmatrix} 354 \\ 251 \\ 327 \\ 328 \\ 301 \\ 270 \\ 330 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_{mean} \\ b_{2000} \\ b_{2001} \\ b_{2002} \end{pmatrix} + \begin{pmatrix} e1 \\ e2 \\ e3 \\ e4 \\ e5 \\ e6 \\ e7 \end{pmatrix}$$

7 x 1 = 7 x 4 4 x 1 + 7 x 1 dimension of matrices

The X matrix contains elements that relate the 7 observations to the effects or attributes we consider in the model. Observation 1 (354Kg) has a full dose of the mean, plus a full dose

of the year-2000 effect. X is an incidence matrix telling us whether a certain effect j (column j) is present or not with respect to a certain observation i (row i).

Solutions for the effects are found using the Least Squares equations:

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

where $^{-1}$ refers to "inverse" and the matrices look like

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

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

A complication is now that there is a dependency (or a redundancy) in the set of equations. The columns of X add up to each other. This is always the case if we have a mean and an additional 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. 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 affect the estimates, but not the relevant comparisons, and in this case, it does not affect the estimate of the year difference.

The second option is the easiest. If a parameter is set to zero you can omit the equation for that parameter. If a certain year has a zero solution, the general mean will be in fact represent the estimate of the mean of that year. The other year effects are deviations/differences from the year that was set to zero,

The first option, i.e. set the general mean to zero, 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 $X'Y$:

$$X' \qquad\qquad X \qquad = \qquad X'X$$

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

Note the incidence matrix. The last year is represented as a function of the two previous years: $b_{2000} + b_{2001} + b_{2002} = 0 \rightarrow b_{2002} = -b_{2000} - b_{2001}$.

On inversion, $(X'X)^{-1} = \begin{pmatrix} 0.1944 & -.0278 & -.1111 \\ -.0278 & 0.3611 & -.0556 \\ -.1111 & -.0556 & 0.2778 \end{pmatrix}$

$$X' \qquad\qquad Y \qquad = \qquad X'Y$$

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

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

$$\begin{pmatrix} b_{mean} \\ b_{2000} \\ b_{2001} \end{pmatrix} = \begin{pmatrix} 0.1944 & -.0278 & -.1111 \\ -.0278 & 0.3611 & -.0556 \\ -.1111 & -.0556 & 0.2778 \end{pmatrix} \begin{pmatrix} 2161 \\ 275 \\ 896 \end{pmatrix} = \begin{pmatrix} 313 \\ -10.5 \\ -6.5 \end{pmatrix}$$

The first solution (313) refers to the estimated mean; other solutions are year effects of Y_{2000} and Y_{2001}. There is no solution for Y_{2002} but this can easily be worked out as $b_{2002} = -b_{2000} - b_{2001} = 10.5 + 6.5 = +17$. Hence the difference between years 2002 and 2000 is 27.5 kg.

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

General mean zero		First year zero (b_{2000}=0)		Last year zero (b_{2002}=0)		Sum of years to zero (b_{2000}+ b_{2001} + $_{2002}$=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	
Y_{2000} = 302.5		Y_{2000}= 0		Y_{2000} = -27.5		Y_{2000} = -10.5	
Y_{2001} = 306.5		Y_{2001} = +4		Y_{2001} = -23.5		Y_{2001} = -6.5	
Y_{2002} = 330		Y_{2002} = +27.5		Y_{2002} = 0		Y_{2002} = 17	

We see from the different restrictions that the important parameters are always the same. These important parameters result from 'estimable functions', those are linear combinations of observations. The expectation of any observation according to our model is

$$E(y_i) = \mu + Y_{2002}$$

And the difference between observations in two different years is

$$E(y_i - y_j) = \mu + Y_i - \mu - Y_j = Y_i - Y_j$$

Therefore, the difference between 2 year effects are estimable, and from the solutions we see that these differences are the same, not affected by the constraint we put on the solutions.

Year differences:
$Y_{2002} - Y_{2000} = 27.5$
$Y_{2002} - Y_{2001} = 23.5$
$Y_{2001} - Y_{2000} = 4$

A year effect by itself is not estimable, we can not find a combination of observation to estimate Y_i as this is always confounded with μ. Another estimable function is the expected value of an observation, i.e. $\mu + Y_i$.

Expected observations: $\mu + Y_{2000} = 302.5$

$$\mu + Y_{2001} = 306.5$$
$$\mu + Y_{2002} = 330$$

and again these values are the same for all sets of solutions

Model with Two Fixed Effects

The previous section might have looked unnecessarily complicated as with only one fixed effect in the model, these year differences can be estimated from the raw means for each year. However, the story is different if we have more than one fixed effect. In that case the means of each fixed effect have to be adjusted for the other effect. In a linear model, we easily add extra effects to the model.

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. In a linear model, joint estimation for several effects is elegant and relatively simple.

The model for two fixed effects becomes:

Weight = general mean + effect of year + effect of sex + random error

$$y_{ijk} = \mu + Y_i + S_j + e_{ijk}$$

In matrix notation:

$$y = Xb + e$$

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 2002 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 2002 the effect of year 2000 (relative to 2002) the effect of year 2001 (relative to 2002) the effect of males (relative to females)

Note that the effect of year 2002 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 2002 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 2002. Notice also that the difference between 2000 and 2001 has not changed. This is because within these years there were equal numbers of males and females.

Correcting for other fixed effects is different from just taking raw means *only* if those other effects are unequally contributing to a fixed effect under consideration, e.g. if not all years have an equal amount of males and females. This is called a "balanced design". In practice we hardly ever have a balanced design, and we need a linear model to correct appropriately for all other effects.

The reasoning for using linear models to disentangle different fixed effects is also true for disentangling genetic and systematic environmental effects. To estimate herd effects, we need to take into account that some herds may have used better sires, and therefore have on average animals with better genetic effects. Jointly estimating fixed systematic environmental effects and random effects of animals' breeding values is accommodated for in a mixed model.

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
```

Model with a Covariate

In the two previous sections, we considered two fixed effects, both of them being class variables. Another type of fixed effect can be due to a continuous variable, and the most obvious example is age at measurement. Suppose the 7 animals were measured at slightly different ages. To correct the phenotypes for an age effect, we can first estimate this age effect in a linear model. A model with just one continuous variable is a regression model. For example, if we fit just age (with the age at measurement measured in months):

Weight = general mean + age of the animal

$y_i = \mu + age_i + e_i$

In matrix notation:

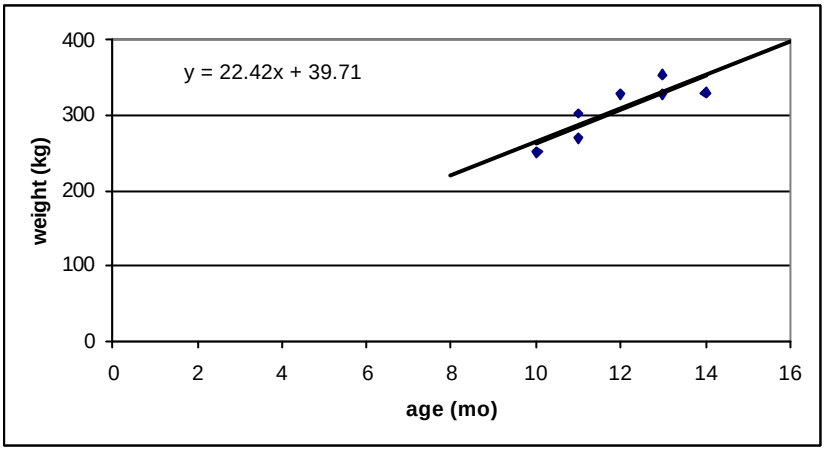
$y = Xb + e$

The X matrix, and the solution become:

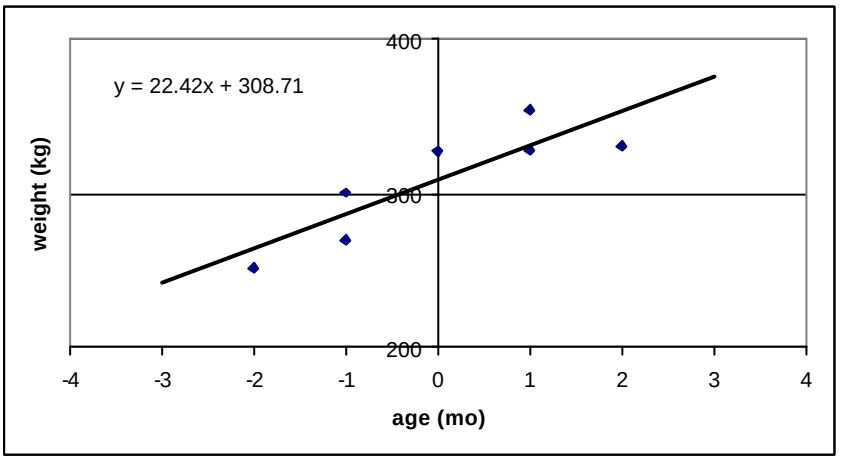
X	$\hat{b}$	meaning
$\begin{pmatrix} 1 & 13 \\ 1 & 10 \\ 1 & 12 \\ 1 & 13 \\ 1 & 11 \\ 1 & 11 \\ 1 & 14 \end{pmatrix}$	$\begin{pmatrix} 39.71 \\ 22.42 \end{pmatrix}$	the intercept: the weight at age 0 the slope: the extra weight for one extra month of age

Hence, there is only one column needed for a covariable: it takes only one degree of freedom, or one parameter to estimate (unless we also want to fit a quadratic effect of age), and rather than an "incidence" we simply put the age of measurement in the X-matrix. In a regression model we estimate the 'mean' as an intercept, which is basically the estimated weight at age 0. In this example, the value is pretty realistic to represent 'birth weight' but it could have been quite deviant from a realistic value as the observed values for age are far from birth weight, so it is based on extrapolation. Again, we can reparameterise the model,

by expressing the covariable relative to a mean. We can choose to express age relative to 12 months, and the estimated intercept will become the estimated weight at 12 months.



Regression analysis with observed ages (above) and ages as deviations (below)



X	$\hat{b}$	meaning
$\begin{pmatrix} 1 & +1 \\ 1 & -2 \\ 1 & 0 \\ 1 & +1 \\ 1 & -1 \\ 1 & 2 \\ 1 & +1 \end{pmatrix}$	$\begin{pmatrix} 308.71 \\ 22.42 \end{pmatrix}$	the intercept: the weight at the age of 12 months the slope: the extra weight for one extra month of age

Hence, reparameterisation affects the solution for the intercept, but not the slope. We are still estimating that the animals grow 22.42 kg per month.

Now, to put it all together, we can fit a model where the effects of year, sex and age are jointly fitted.

$$y_{iik} = \mu + Y_i + S_j + age_{ijk} + e_{ijk}$$

In matrix notation: $y = Xb + e$

The X matrix, and the solution become:

X	$\hat{b}$	meaning
$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & -2 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & -1 \\ 1 & 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 250.5 \\ 51.8 \\ 49.2 \\ 26.7 \\ 26.4 \end{pmatrix}$	The mean for 2002, females, 12 mo The effect of 2000 relative to 2002 The effect of 2001 relative to 2002 The effect of males relative to females The effect of one extra month of age

Now, after correcting all effects for each other we observe that

- The difference between Y_{2000} and Y_{2001} is a bit smaller than 4, because the ages of the animals in those years were different.
- The difference between Y_{2000} and Y_{2001} on one hand, and Y_{2002} on the other hand have changed drastically. Because animal 7 was measured at an older age, the year effect of 2002 has gone down.
- The difference between males and females is also reduced because differences between males and females are partly explained by age of measurement (males were on average 12.5 months of age and females were on average 11.3 months old; animal number 7 does not count as he could not be compared with a female within his year).

Model with interaction

Consider the following small data set with 8 observations on weight from 2 breeds (Angus and Brahman) in two environments (Tropical and Temperate):

```
1 BRA TROP 255
2 BRA TROP 245
3 ANG TROP 262
4 ANG TROP 238
5 BRA TEMP 295
6 BRA TEMP 305
7 ANG TEMP 345
8 ANG TEMP 355
```

A simple 2 way analysis:

```
analysis of test data 6 LM course
  ID
  breed 2 !A
  environm 2 !A
  weight
exmp6.dat
weight  ~ mu breed environm
  environm                TROP                  0.000       0.000
  environm                TEMP                  75.00       12.99
```

```
  breed                    BRA                    0.000       0.000
  breed                    ANG                    25.00       12.99
  mu                             1                237.5       11.25


Analysis of Variance                  DF     F-incr
   5 mu                                1    1958.68
   2 breed                             1       3.70
   3 environm                          1      33.32
```

However, it is useful to make a Table with the sub class means for the different combinations, like this

	TROP	TEMP	
BRA	250	250	250
ANG	300	350	325
	275	300	287.5

We see that there is definitely a breed effect in the temperate environment, but not in the tropics. This example calls for a test about whether there is an *INTERACTION* between the main effects. Practically, to test whether one effect depends on the levels of another effect, i.e. whether the breed effect depends on whether you are in the tropics or in Scandinavia.

An interaction model: The model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + ?_{ij} + e_{ij},$$

where $?_{ij}$ is an interaction term. The number of levels for ? is equal to the number of filled subclasses. With no missing data, that would be equal to nr. of levels of effect 1 x nr. of levels for effect 2.

```
analysis of test data 6 LM course
  ID
  breed 2 !A
  environm 2 !A
  weight
exmp6.dat
weight  ~ mu breed environm breed.environm
```

The X matrix look like (parameters printed above)

$$\begin{pmatrix} m & a_1 & a_2 & b_1 & b_2 & g_{11} & g_{12} & g_{21} & g_{22} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The incidence matrix has many dependencies; all main effects add up to the mean, as before, but now also, all interactions within a main effect subclass add up to that main effect. This means technically that it is hard to estimate both main effects and interactions. We can set restriction to the solutions, as before, but the interpretation becomes a bit more tricky. See below for two possible solutions, and try to interpret these, with an eye on the means table above. You can imagine that with missing subcells and with more levels, such interpretation could become more difficult (that's why it is usually good to try a very simple example first.
with solutions:

```
not fitting μ

  breed.environm       BRA.TROP                  0.000       0.000
  breed.environm       BRA.TEMP                  0.000       0.000
  breed.environm       ANG.TROP                  0.000       0.000
  breed.environm       ANG.TEMP                  50.00       14.80
  environm                 TROP                  0.000       0.000
  environm                 TEMP                  50.00       10.46
  breed                    BRA                   250.0       7.399
  breed                    ANG                   250.0       7.399

fitting μ and setting first of levels to 0.

  breed.environm       BRA.TROP                  0.000       0.000
  breed.environm       BRA.TEMP                  0.000       0.000
  breed.environm       ANG.TROP                  0.000       0.000
  breed.environm       ANG.TEMP                  50.00       14.80
  environm                 TROP                  0.000       0.000
  environm                 TEMP                  50.00       10.46
  breed                    BRA                   0.000       0.000
  breed                    ANG                   0.000       10.46
  mu                              1              250.0       7.399
```

and analysis of variance:

```
Analysis of Variance                  DF     F-incr
   5 mu                                1    6038.81
   2 breed                             1      11.42
   3 environm                          1     102.74
   6 breed.environm                    1      11.42
```

We will discuss interactions further in a next chapter on analysis of variance.

For the moment, we are mainly interested in the meaning of the solutions obtained. The question here is: how relevant are the estimates of main effects in a dataset where significant interactions exist.

RECOMMENDED (backup) READING

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

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

Searle, S.R. 1971. Linear Models. Wiley & Sons.

Exercise 3: Linear models

A) Example of regression on covariables variables

Give are the weights of six cows, and their age at which the measurement was taken

cow	ages(mo)	weight (kg)
1	18	494
2	21	556
3	19	542
4	17	473
5	23	632
6	19	544

Determine the effect of age on weight

Determine the first order regression
Compare the residuals

Hint: The model is: y = Xb + e. Determine X and y for this example (see Ch.2)
Solutions are via Least squares: $\hat{b}$ = (X'X)$^{-1}$X'y
Residuals calculated as $\hat{e} = y - X\hat{b}$. Residual sums of squares = SSE = $\hat{e}'\hat{e}$
Residual variance = $\hat{e}'\hat{e}$/(n-2)

B) Example of regression on class variables

Given are the weights of the same cows, but now we have discovered they are from 2 different breeds, and they were raised under two different feeding regimes (intensive and extensive grazing)

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

Determine the effects of breed and feeding regime by Least Squares Analysis
(You may first assume equal ages now, but then also try to account for it)

Compare estimates from different models and explain.