School of Mathematics, Statistics and Computer Science

STAT404 - Mixed Models Lecture Notes

R. D. Murison

Printed at the University of New England, February 18, 2008

Coordinator

Lecturer: A Prof Bob Murison
Address: School of Mathematics, Statistics and Computer Science
University of New England, Armidale, NSW, 2351.
Telephone: 02 6773 2544
Fax: 02 6773 3312
Email : rmurison@turing.une.edu.au
Web site : `http://turing.une.edu.au/~rmurison`

Objectives

To extend the linear model to incorporate random effects in several ways.

Methods that use the Bayesian paradigm will be added later.

Reference Books

These are listed in the references of the notes.

- Mixed effects models in S and S-PLUS. Pinheiro and Bates.
- Data Analysis Using Regression and Multilevel/Hierarchial Models. Gelman and Hill.
- Analysis of Longitudinal Data. Diggle, Liang and Zeger.
- Semiparametric Regression. Ruppert, Wand and Carroll.
- Nonparametric Regression and Generalized Linear Models. Green and Silverman.

Assessment

Assignments will count 50% towards the total assessment. The other 50% will be by a take-home examination at the end of First Semester. To obtain a pass, students are expected to perform at least close to pass standard in each of the two sections.

Assignments are to be of a high standard with neat setting out, appropriate captioned Tables and Figures of results and descriptions of the analysis.

UNE policy on plagiarism

I draw your attention to UNE's policy on plagiarism stated at
`http://www.une.edu.au/policies/pdf/plagiarismcoursework.pdf`

By this stage of your degree, I am sure that this is understood but we operate on the understanding that you are conversant with the policy.

Contents

Part I

Unit Notes

Chapter 1

The Linear Mixed Model and the Randomized Block Design

The statistical model consists of fixed and random components. The fixed effects are represented by low dimension expressions such as curves, means etc and the random components by densities or probability fuctions.

Example 1.1

An ecologist wanted to measure weed invasion into a woodland from a clearing. He set up 4 transects with sample points at $0, 5, 10, 20, 30, 50$ metres away from the edge to monitor the establishment over 2 years of a number of species. Colonization depends upon seed dispersal, season etc and for different species, the the expected profiles are shown in Figure 1.1.

Apart from the first case (known as ubiquitous), a non-linear profile was anticipated. Would a different design with 2 transects but with closer sampling within a transect be better? That consideration depends on the primary *interest*, whether it is to

(a) model the profile across the woodland, or

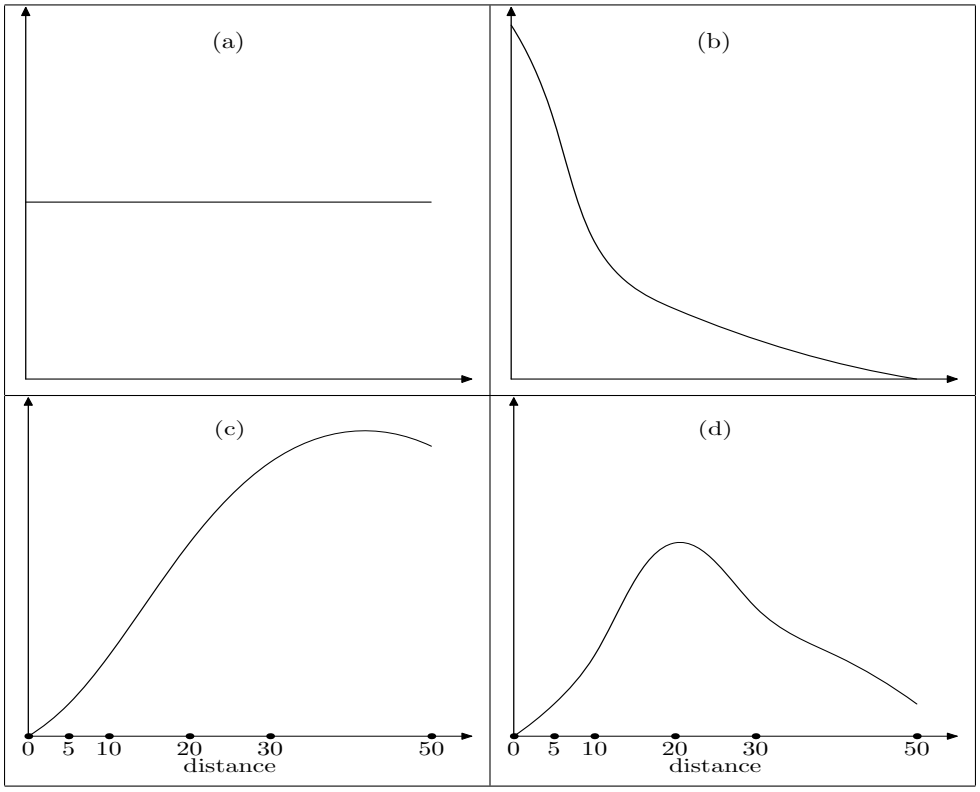
(b) model the variability or clustering of the weed.

When we consider how to allocate sampling points to an experiment design, we need to make choices from the following table,

	Fixed	Random
Interest		
Nuisance		

In (a), the fixed effects are interest and the random effects nuisance whereas in (b) the interest is in the random effects. Designs are seldom optimal in all facets but do not have to be inadequate in secondary interests.

Figure 1.1: Anticipated weed counts across a woodland



1.1 Randomized Block

Example 1.2

This example is taken from Pinheiro and Bates [7]. The data are measures of effort required by 9 Subjects to arise from each of 4 stool Types and the design is a randomized block with each subject constituting a block.

	Subject									
Type	1	2	3	4	5	6	7	8	9	
1	12	10	7	7	8	9	8	7	9	**8.6**
2	15	14	14	11	11	11	12	11	13	**12.4**
3	12	13	13	10	8	11	12	8	10	**10.8**
4	10	12	9	9	7	10	11	7	8	**9.2**
	12.25	**12.25**	**10.75**	**9.25**	**8.5**	**10.25**	**10.75**	**8.25**	**10.0**	

The R code to find these data is

```
library(nlme)
data(ergoStool)
```

A basic linear model has only 1 random effect, the experiment error. In the above data set, Subjects are a random sample from a population and are random effects so the model has 2 random effects- subjects and error. The terminology *mixed models* is used when there are models for the fixed effects and more than 1 random effect but the generic term *linear model* is consistent for models where the fixed and random components are additive.

For the vector of responses for Subject i,

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \epsilon_i \\ b_i \sim N(0,\sigma_b^2), \quad \epsilon_i &\sim N(0,\sigma^2 I) \quad , \mathrm{cov}(b,\epsilon) = 0 \end{aligned} \tag{1.1}$$

where

$$X_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad , Z_i = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{1.2}$$

The Z matrix is an matrix of 1's or zeroes to indicating subject i.

Over the whole data set,

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_9 \end{bmatrix}_{36,1} = \begin{bmatrix} X_1 \\ \vdots \\ \vdots \\ X_9 \end{bmatrix}_{36,4} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}_{4,1} + \begin{bmatrix} Z_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & Z_9 \end{bmatrix}_{36,9} \begin{bmatrix} b_1 \\ \vdots \\ b_9 \end{bmatrix}_{9,1} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_{36} \end{bmatrix}_{36,1}$$

The contrasts used in equation (1.2) are treatment contrasts leading to the following interpretations of the parameters,

β_1	mean of stool 1
β_2	effect of stool 2 compared to stool 1
β_3	effect of stool 3 compared to stool 1
β_4	effect of stool 4 compared to stool 1

Other contrasts are simply different ways of spanning the parameter space and one set of linearly independent contrasts can be transformed into another.

The analysis is done vie the lme() function and so continuing on from the previous code,

```
mmodel <- lme(effort ~ Type,random=~1 | Subject,data=ergoStool)
print(anova(mmodel))
print(summary(mmodel))
b <- random.effects(mmodel)    # or ranef(mmodel)
```

In the above code, you can see separate models for fixed and random components.

The estimation procedure for linear models with random effects is REML which stands for Residual Maximum Likelihood. For balanced data (like these), the results are the same if Subjects (or blocks) are fitted as a fixed effect. However, blocks are nuisance effects and we do not want to restrict the analysis by requiring interest effects be balanced over nuisance effects. For unbalanced data, the results from

```
lme(effort ~ Type,random=~1 | Subject)
```

and

```
lm(effort ~ Type + Subject)
```

would differ.

The usual caveats apply; accept the model if the assumptions are not contravened. To finish off this job, you would need to check the residuals,

```
plot(mmodel)
```

A newer version of `lme` is `lmer` which is contained in the package `lme4`. This is widely used in the text "Data Analysis Using Regression and Multilevel/Hierarchial Models" by Andrew Gelman and Jennifer Hill [3]. They augment `lme4` with other functions contained in the package `arm` which is companion to the textbook. An example of these is the function `display`.

```
library(lme4)
mmodel2 <- lmer(effort ~ Type + (1|Subject),data=ergoStool)
print(anova(mmodel2))
```

Gelman and Hill describe models such as these as *varying-intercept* models and this model is multi-level in the sense that the relationship amongst types varies across subjects.

Example 1.3

These data plotted in Figure 1.2 are productivity scores for each of 6 randomly chosen workers tested on each of 3 different machine types. Each worker used each machine 3 times - ie. 3 replicates.

We observe that variability due to replication is of lower order of magnitude than variability amongst workers. Further exploratory data analysis (Figure 1.3) suggests a Worker × Machine interaction.

A model whose random effects is worker (only) would be

$$\begin{aligned} y &= X\beta + Z_1 b + \epsilon \\ b \sim N(0, \sigma_b^2 I) &\quad , \quad \epsilon \sim N(0, \sigma^2 I) \end{aligned} \tag{1.3}$$

implemented by

```
machine1 <- lme(score ~ machine,random=~1 | Worker,data=Machines)
```

A straightforward extension of equation (1.3) to include the obvious Worker × Machine interaction, gives

$$\begin{aligned} y &= X\beta + Z_1 b + Z_2 m + \epsilon \\ b \sim N(0, \sigma_b^2 I) \;\; , m &\sim N(0, \sigma_m^2 I) \quad , \epsilon \sim N(0, \sigma^2 I) \end{aligned} \tag{1.4}$$

which is fitted by

Figure 1.2: Worker productivity on machines

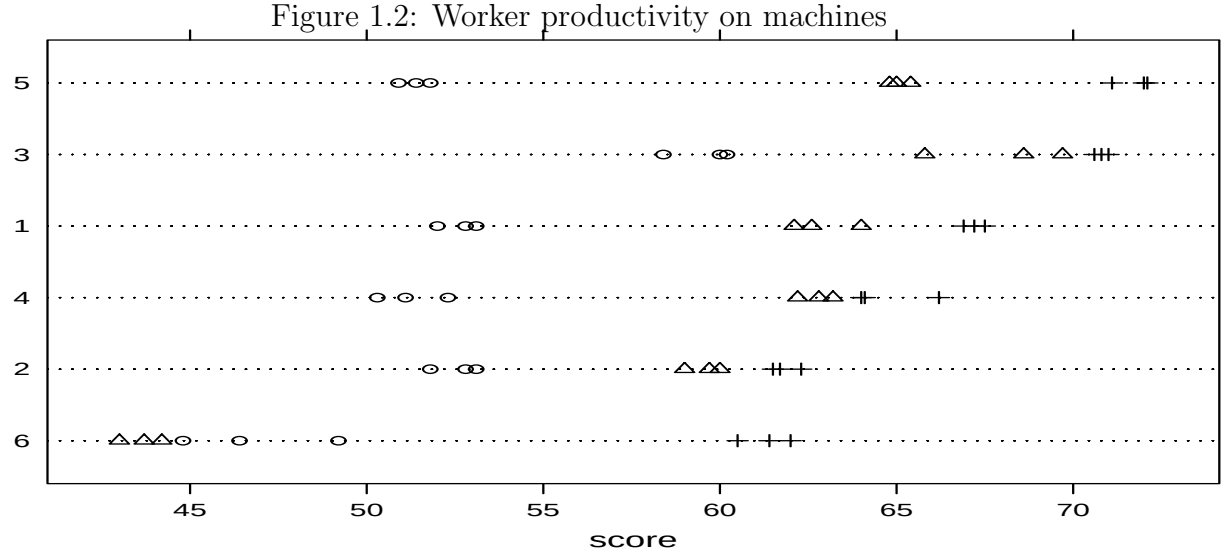
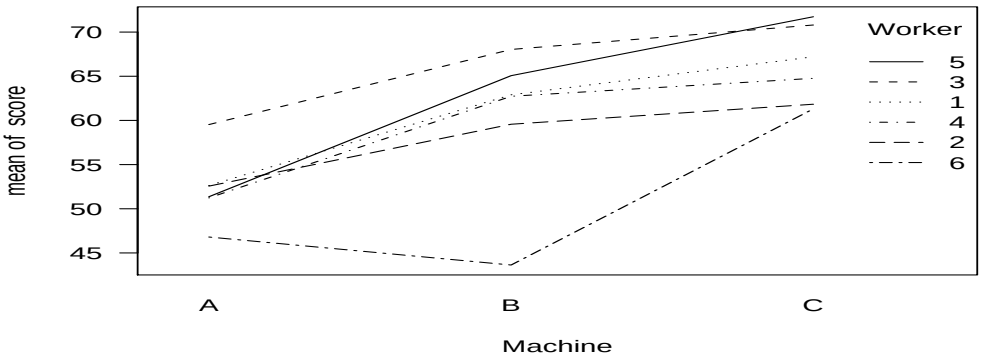


Figure 1.3: Worker × machines interaction



```
machine2 <- lme(score ~ machine,random=~1 | Worker/Machine,data=Machines)
```

These models are nested so we can compare them by a likelihood ratio test,

```
anova(machine1,machine2)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
machine1     1  5 296.8782 306.5373 -143.4391
machine2     2  6 227.6876 239.2785 -107.8438 1 vs 2 71.19063  <.0001
```

In the model given at (1.4), the random interactions have the same variance, σ_m^2 and the repeated measures from each subject are assumed independent. A more general model is

$$\begin{aligned} y_i &= X_i\beta + Z_i b_i + \epsilon_i \quad i = 1, \ldots, 6 \\ b \sim N(0, \psi) &\quad , \quad \epsilon \sim N(0, \sigma^2 I) \end{aligned} \tag{1.5}$$

where, ψ is a positive-definite symmetric 3×3 matrix.

The code for this is implemented by

```
machine3 <- lme(score ~ machine,random=~Machine -1 | Worker,data=Machines)
anova(machine1,machine2,machine3)

         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
machine1     1  5 296.8782 306.5373 -143.4391
machine2     2  6 227.6876 239.2785 -107.8438 1 vs 2 71.19063  <.0001
machine3     3 10 228.3112 247.6295 -104.1556 2 vs 3  7.37635  0.1173
```

So long as we have prescribed the factors correctly, the program generates the design matrices X and Z and they can be extracted for further numerical work if required. For example, the design matrices for Worker 1 are

```
X1 <- model.matrix(score ~ Machine,data=Machines[Machines$Worker==1,])
Z1 <- model.matrix( ~ Machine - 1,data=Machines[Machines$Worker==1,])
print(cbind(X1,Z1))
```

unit	X_1			Z_1		
	(Intercept)	MachineB	MachineC	MachineA	MachineB	MachineC
1	1	0	0	1	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
19	1	1	0	0	1	0
20	1	1	0	0	1	0
21	1	1	0	0	1	0
37	1	0	1	0	0	1
38	1	0	1	0	0	1
39	1	0	1	0	0	1

Note that the columns of Z_1 span the space of X_1 but the strategy of REML estimation does not have these matrices in conflict because the analysis projects the data into the orthogonal spaces of (i) treatment contrasts and (ii) error contrasts by

$$\begin{bmatrix} L_1^T \\ L_2^T \end{bmatrix} y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

where L_1, L_2 satisfy $L_1^T X = X_p$ and $L_2^T X = 0_{(n-p)}$, for example $L_1 = X(X^T X)^{-1} X^T$ and $L_2 = I - X(X^T X)^{-1} X^T$. Then

$$\ell(y) = \underbrace{\ell(y_1)}_{\text{fixed}} + \underbrace{\ell(y_2)}_{\text{random}}$$

We pursue this theory (see Verbyla (1990)) [13] in detail later on.

Chapter 2

Balanced incomplete blocks

Incomplete blocks comprise a set of designs where it is not possible to allocate every treatment in each block. This may arise if there is insufficient homogeneous material to which the treatments are to be applied. A randomized block, where the blocks are known to be heterogeneous, will lead to over inflated experiment error which of course reduces the power of the block.

1. In taste testing, the palate fatigues so it is advisable to restrict the number of samples. If 10 treatments need to be compared and each taster (block) is reliable for only 5, then the treatments need to be allocated so that all 10 treatments can be compared. were we to test only 5 at a time, we would incur the extra variance due to different panels.

2. A physiotherapist is researching the use of the web for remote treatment. The experiment requires that practioners rate 15 conditions by video. A randomized complete block would require each practitioner to examine 15 videos, each of about 1 hour duration. The fatigue factor suggests extra variance due to the order of the assessment (ie. first or last etc.) and the possibility of raters not completing their assessments put the viability of the experiment at risk. The problem is averted with an incomplete block design.

3. In a factory experiment, different factorial combinations are are to be trialled. But as the experiment progresses, the environment is likely to change as the factory warms up so that the treatments are measured under different ambient conditions. To get fair comparisons, block size should be restricted so that treatments are measured under homogeneous conditions.

4. In a 20 team competition, there is not enough time for each team to play the others twice in a season. After round 1, the competition draw becomes an incomplete block experiment which allows favourable comparison of points.

From (10.6), $\text{var}\hat{\beta} = \sigma^2(X^TX)^{-1}$. The optimal variance of the parameters is a balance between the degrees of freedom for the design matrix X and the variance, σ^2. A large

RCB may reduce $|(X^TX)^{-1}|$ but at the expense of σ^2. Small block sizes will ameliorate this and the right balance is sought. This balance may be influenced as much by practical experiment reasons as by the mathematics.

In chapter 1, it was mentioned that in a RCB (ie. each treatment occurs the same number of times across blocks), blocks and treatments are orthogonal so blocks could be fitted as fixed effects and Total SS = Block SS + Treat SS. However the mixed model and REML estimation does not demand balance because the model has $\text{cov}(\epsilon, b) = 0$ and the estimation is via a fixed model and a random model which are orthogonal. Hence the balance restriction is freed by REML and since Incomplete blocks do not have each treatment occurring in each block an equal number of times, the convenient analysis is via the mixed model.

A major class of non–orthogonal designs are those known as **balanced incomplete block** designs. As the word *incomplete* suggests, not all treatments occur in each block. The *balance* referred to is a general balance (i.e. it refers to such matters as the number of times each pair of treatments occur together in a block) rather than meaning all treatments occur together in all blocks.

2.1 Balanced incomplete blocks

When systematic differences exist between units, blocking is often used as a device to improve precision. The blocks are formed from groups of similar units, and if all the treatments under investigation are applied randomly within each block, then it is possible to make a fair comparison between the treatments. Unfortunately the blocks may not always contain enough units to accommodate all the treatments and an alternative solution to Randomised Blocks is needed if the number of treatments is not to be reduced. If all treatment comparisons are equally important then the most satisfactory design is *Balanced Incomplete Blocks* (BIB).

The BIB design has three main properties

(a) all blocks have the same number of units,

(b) all treatments are equally replicated,

(c) all treatment pairs occur in the same block equally often over the whole design.

It is convenient to introduce a standard notation to describe the features of a BIB design, and these parameters are now widely accepted.

- b number of blocks
- v number of treatments (sometimes t)
- k number of units in a block
- r number of replicates for each treatment
- n total number of units
- λ concurrence

The concurrence is the number of times two treatments occur together in the same block. The properties of a BIB design may therefore be stated in the alternative form that $v > k$ and k, r, λ are constant over all blocks and treatments.

Example 2.1

A typical example of a BIB design is given below.

	Blocks									
units	1	2	3	4	5	6	7	8	9	10
(i)	1	1	1	1	1	1	2	2	2	3
(ii)	2	2	2	3	3	4	3	3	4	4
(iii)	3	4	5	4	5	5	4	5	5	5

Observe that treatments are conventionally described by numbers. The above plan is in its basic form before randomisation. In practice, the experimenter should allocate the numbers randomly to the treatments, randomise the block order and the assignment of the selected treatments to the units within each block. In the example given it is seen that $b = 10, v = 5, k = 3, r = 6, n = 30, \lambda = 3$.

If the rows are considered as blocks the design is a BIB with $b = v = 7$ and $k = r = 4$ and $\lambda = 2$. Notice however that each column contains each treatment exactly once, so this type of design is suitable when blocking is required in two directions. The double blocking is reminiscent of the properties of a Latin Square and this type of design is called a *Youden Rectangle*. An alternative description for Youden Rectangles is Incomplete Latin Squares, and a method of generating them is to omit certain columns of particular Latin Square designs. It is not always possible to find a Youden Rectangle of a prescribed specification, but it is always possible to find a Youden rectangle where the number of treatments is one more than the block size. All that is needed is for the last column to be dropped from the appropriate Latin Square and a Youden Rectangle will always result. However it is not so easy if more than two columns need be dropped.

For the particular design in **Example ??** the complete Latin Square is formed through augmentation with a further Youden rectangle.

1	2	3	6	4	5	7
2	3	4	7	5	6	1
3	4	5	1	6	7	2
4	5	6	2	7	1	3
5	6	7	3	1	2	4
6	7	1	4	2	3	5
7	1	2	5	3	4	6

The advantage of a resolvable BIB is that it can also be analysed as a RCB if the blocks are not significantly different (or that they do not account for sufficient extraneous variation). The extra degrees of freedom can be pooled into a single random term, ie the error, and we may improve the power of the design to detect differences by increasing error df without inflating the variance. But we may make big inroads into the error by removing block effects.

We can compare 2 designs by the ratio of the variance of a treatment contrast from design 1 to its counterpart of design 2 and this is termed EFFICIENCY. The efficiency of a BIB to a RCB is given by,

$$\begin{aligned} E &= \frac{\text{var}(\tau_i - \tau_j)_{BIB}}{\text{var}(\tau_i - \tau_j)_{RCB}} \\ &= \frac{2\sigma^2/r}{2k\sigma^2/\lambda v} \\ &= \lambda v/rk = (1-1/k)/(1-1/v) \leq 1 \end{aligned} \tag{2.1}$$

Certain resolvable designs have arisen to maximise efficiency compared to a RCB,

- $v = b^2$, $k = b$, is called a square lattice,
- $v = b(b-1)$, $k = b-1$ is called a rectangular lattice,
- $v = b(b-l)$, $k = (b-l)$ is called an alpha-design

2.2 Statistical model for BIB designs.

The model is similar to the randomized block design,

$$y_{ij} = \tau_i + b_j + \epsilon_{ij} \tag{2.2}$$

but each block contains only a subset of the treatments.

If we treat blocks as fixed effects and remove their effects prior to examining treatments, the information about the treatments comes only from comparisons within blocks and is known as an intra-block analysis. However, we should treat blocks as another random component because they are random samples from a population so the random parts of the model (2.2) are modelled by

$$\text{inter-block, } b_j \sim N(0, \sigma_\beta^2) \tag{2.3}$$

$$\text{intra-block, } \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \tag{2.4}$$

$$E(Y_{ij}) = \tau_j \tag{2.5}$$

$$\text{var}(Y_{ij}) = \sigma_b^2 + \sigma_\epsilon^2 \tag{2.6}$$

Maximum likelihood estimates can be derived from

$$\ell(y; \tau, \sigma_\beta^2, \sigma_\epsilon^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_\beta^2 + \sigma_\epsilon^2) - \frac{1}{2}\sum_i \sum_j \frac{(y_{ij} - \tau_i)^2}{(\sigma_\beta^2 + \sigma_\epsilon^2)}$$

2.3 The old way

For *balanced* incomplete blocks, the appropriate variances can be determined by AOV.

Source	df	MS	E(MS)
Treat	(v-1)	MS(Treat)	$\sigma^2_\epsilon + \frac{(kb-b)}{(v-1)} \sum_i \tau^2_i$
Blocks	(b-1)	MS(Blocks)	$\sigma^2_\epsilon + \frac{(kb-v)}{(b-1)} \sigma^2_\beta$
Error	bk-v-b-1	MS(Error)	σ^2_ϵ
Total	(bk-1)		

Thus for the balanced case, equating expected and observed mean squares leads to estimates of the variance components and better estimates of the variances of treatment means, eg (2.6). This is termed an inter-block analysis.

Each block is intended to be relatively homogeneous and so intra-block variance should be less than inter-block variance.

If the treatments are not balanced across blocks, the AOV method will not work because we cannot plug in the block size k. The REML estimation arose from this situation, (Patterson & Thompson, 1971), but rapidly found widespread application in all facets of statistics.

Example 2.2

Below is a set of data whose design is a balanced incomplete block, followed by an old-fashioned analysis.

```
blocks treat response
1      A     20
1      B     23
1      C     16
2      A     22
2      C     19
2      D     17
3      B     25
3      C     20
3      D     21
4      A     16
4      B     20
4      D     14
```

```
lmodel <- lm(y ~ treat + blocks)
print(anova(lmodel))

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
treat      3   48.2    16.1    13.6 0.0077 **
blocks     3   58.8    19.6    16.6 0.0050 **
Residuals  5    5.9     1.2
```

If the blocks term was not significant, we would resort to a completely random design. However, it is significant here and because of non-orthogonality there is some information about the treatments contained in the blocks.

To get estimates of the treatment means, we get a prediction of what each treatment would have been if it had been allocated to each block. This is done by using the linear model. In the above example, the block and treatments effects are given by

```
BIB.effects <- summary(BIB.lm)$coef                    #$

             Value Std. Error t value  Pr(>|t|)
(Intercept)  19.792     0.8308  23.822 2.428e-06
    blocks2   1.625     0.9421   1.725 1.451e-01
    blocks3   3.375     0.9421   3.583 1.583e-02
    blocks4  -3.000     0.9421  -3.184 2.442e-02
     treatB   2.750     0.9421   2.919 3.305e-02
     treatC  -3.125     0.9421  -3.317 2.107e-02
     treatD  -3.125     0.9421  -3.317 2.107e-02
```

The predictions of treatments means for each block are shown in Table 2.1. We know that treatment A was not in block 3 but we get an estimate of it as if it did occur there. Likewise treatment B in block 2, treatment C in block 4 and treatment D in block 1 are estimated.

With less effort, we can see that the estimates of treatment means for A, B, C, D are

$$\begin{bmatrix} \hat{\mu_A} \\ \hat{\mu_B} \\ \hat{\mu_C} \\ \hat{\mu_D} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 1 & 0 & 0 \\ 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 1 & 0 \\ 1 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 19.79 \\ 1.625 \\ 3.375 \\ -3 \\ 2.75 \\ -3.125 \\ -3.125 \end{bmatrix}$$

For this experiment $b = 4$, $k = 3$, $v = 4$ so equating the observed Mean Squares in the AOV to their expected values yields,

$$\sigma^2_\beta = \frac{(19.6 - 1.2)}{(3 \times 4 - 4)/(4 - 1)} = 6.9$$

Table 2.1: Estimates of treatment means for each block of a BIB

Block	Treat	Intercept		block eff		treat eff		treat est	mean
1	A	19.79					=	19.79	
2	A	19.79	+	1.625			=	21.42	
3	A	19.79	+	3.375			=	23.17	
4	A	19.79	-	3			=	16.79	20.29
1	B	19.79			+	2.75	=	22.54	
2	B	19.79	+	1.625	+	2.75	=	24.17	
3	B	19.79	+	3.375	+	2.75	=	25.92	
4	B	19.79	-	3	+	2.75	=	19.54	23.04
1	C	19.79			-	3.125	=	16.67	
2	C	19.79	+	1.625	-	3.125	=	18.29	
3	C	19.79	+	3.375	-	3.125	=	20.04	
4	C	19.79	-	3	-	3.125	=	13.67	17.17
1	D	19.79			-	3.125	=	16.67	
2	D	19.79	+	1.625	-	3.125	=	18.29	
3	D	19.79	+	3.375	-	3.125	=	20.04	
4	D	19.79	-	3	-	3.125	=	13.67	17.17

2.4 Analysis using a mixed model

By modelling the block effects as nuisance effects orthogonal to the treatment efects, we simplify the analysis.

```
library(nlme)
blocks <- factor(rep(1:4,rep(3,4)))
treat <- factor(LETTERS[c(1:3,1,3,4,2:4,1,2,4)])
response <- c(20,23,16,22,19,17,25,20,21,16,20,14)

bib.mm <- lme(response ~ treat,random=~1| blocks)
print(anova(bib.mm))

            numDF denDF F-value p-value
(Intercept)     1     5     207  <.0001
treat           3     5      18  0.0043
Linear mixed-effects model fit by REML
 Data: NULL
  AIC BIC logLik
   49  49    -18

Random effects:
 Formula: ~1 | blocks
        (Intercept) Residual
```

```
StdDev:         2.6      1.1

Fixed effects: response ~ treat
            Value Std.Error DF t-value p-value
(Intercept)  20.2      1.47  5    13.8  <.0001
treatB        2.8      0.94  5     3.0   0.031
treatC       -3.0      0.94  5    -3.2   0.024
treatD       -3.1      0.94  5    -3.3   0.022
```

To get the treatment means, fit the model again without the intercept.

```
bib.mm2 <- lme(response ~ treat - 1,random=~1| blocks)
print(summary(bib.mm2))

Fixed effects: response ~ treat - 1
       Value Std.Error
treatA  20.2      1.47
treatB  23.0      1.47
treatC  17.2      1.47
treatD  17.2      1.47
```

The variances are recovered with the `VarCorr()` function,

```
print(VarCorr(bib.mm))

            Variance StdDev
(Intercept) 6.900    2.627
Residual    1.183    1.088
```

2.5 Designing an Incomplete Block

GENDEX written by Nam-Ky Nyguen in java, see
`http://designcomputing.hypermart.net/`. has a function called BIB. On turing, type BIB and enter the blocking parameters. Likewise, you can generate ALPHA designs.

Exercise 2.1

Generate a BIB design for $v = 15$, $k = 5$, $b = 9$, $r = 3$.

(I propose that you do this using GENDEX which will require you to login to Turing. Contact me for assistance. Bob Murison)

Simulate data for $b \sim N(0, 10)$ and treatment means $10, 10, 10, 8, 8, 8, 6, 6, 6, 5, 5, 5, 8, 8, 8$ and analyse the simulated data using a mixed model. Plot a histogram of the random block effects and tables the treatment estimates.

Chapter 3

Split plot designs

3.1 The model

The model for a randomised block design arranged in split plots is

$$y_{ijk} = \underbrace{\mu + \theta_i + \beta_j + \epsilon_{ij}}_{\text{main plots}} + \underbrace{\tau_k + (\theta\tau)_{ik} + \delta_{ijk}}_{\text{sub plots}}$$

where β_j is a block effect, θ_i is a main treatment effect and μ_k is a sub-plot treatment effect. It is important to notice that there are two error terms ϵ and δ. Both errors are assumed to be independent $\epsilon \sim N(0, \sigma_m^2)$ and $\delta \sim N(0, \sigma^2)$ and the blocks are random factors, $\beta_j \sim N(0, \sigma_b^2)$

In the following section, b represents block size and m is the number of main plots.

We have from the linear model the following variances

$$\begin{aligned}
\text{var}(Y_{ijk}) &= \sigma_m^2 + \sigma^2 \\
\text{var}(\bar{Y}_{i\bar{i}k}) &= \frac{\sigma_m^2 + \sigma^2}{b} \\
\text{var}(\bar{Y}_{\bar{k}}) &= \frac{\sigma_m^2 + \sigma^2}{bm} \\
\text{cov}(Y_{ijk}, Y_{ijl}) &= \frac{\sigma^2}{\sigma^2 + \sigma_m^2}
\end{aligned}$$

The model formula for covariance amongst sub-plot means is all important in designing and interpreting split-plot designs. We see that if the split-model holds, there is a positive and constant correlation between sub-plot means. If this does not hold, the model is awry and the design is a dud.

Table 3.1: Expected Mean Squares for A split experiment

Source	df	E(MS)
Blocks	(b-1)	
Main	(m-1)	$\sigma^2 + t\sigma_m^2 + \frac{tb}{(m-1)}\sum \mu_i^2$
B×M (Error a)	(b-1)(m-1)	$\sigma^2 + t\sigma_m^2$
Treat	(t-1)	$\sigma^2 + \frac{mb}{(t-1)}\sum \tau_k^2$
M×T	(m-1)(t-1)	$\sigma^2 + \frac{b}{(m-1)(t-1)}\sum(\mu\tau)_{ik}^2$
(Error b)	$(b-1)\left[(t-1)+(m-1)(t-1)\right]$	σ^2
Total	(bmt-1)	

The following excerpt is from the Splus manual, ch 12.

Split-plots are also encountered because of restricted randomization. For example, an experiment involving oven temperature and baking time will probably not randomize the oven temperature totally; but rather only change the temperature after all of the runs for that temperature have been made. This type of design is often mistakenly analysed as if there were no restrictions on the randomisation.

Example 3.1

Five varieties of spring wheat were sown in a randomised blocks design in four blocks. The soil was treated with three different levels of nitrogen randomly allocated to equal areas within each plot. The design and yields in tons/ha were as given.

	V2	V5	V1	V4	V3
	N1 N3 N2	N2 N3 N1	N1 N2 N3	N1 N3 N2	N2 N1 N3
Block 1	4.6 5.5 5.3	5.0 5.4 4.7	5.5 6.1 6.4	5.0 6.0 5.7	5.5 4.9 5.8
	V1	V3	V2	V5	V4
	N3 N1 N2	N1 N3 N2	N3 N2 N1	N2 N3 N1	N2 N1 N3
Block 2	5.8 5.0 5.5	4.9 5.5 5.4	5.4 5.0 4.7	4.6 5.0 4.2	6.2 5.7 6.5
	V5	V1	V3	V2	V4
	N2 N3 N1	N2 N3 N1	N3 N1 N2	N1 N3 N2	N1 N3 N2
Block 3	4.8 5.0 4.6	5.4 5.9 5.0	5.5 4.8 4.7	5.0 5.8 5.1	5.3 6.7 5.8
	V2	V3	V4	V1	V5
	N3 N1 N2	N3 N2 N1	N2 N1 N3	N1 N2 N3	N3 N2 N1
Block 4	5.9 5.0 5.6	4.8 4.6 4.0	5.1 4.7 5.4	5.2 5.5 5.8	5.2 4.8 4.4

The treatment levels were

V1 Timmo	N1 30 kg/ha
V2 Sicco	N2 60 kg/ha
V3 Sappo	N3 90 kg/ha
V4 Highbury	
V5 Maris Dove	

The split plot AOV is:-

```
Error: mainplot
          Df Sum of Sq Mean Sq F Value  Pr(F)
block      3     1.005   0.335   1.047 0.4073
variety    4     6.449   1.612   5.040 0.0128
Residuals 12     3.839   0.320

Error: Within
                 Df Sum of Sq Mean Sq F Value  Pr(F)
nitrogen          2     6.487   3.244   147.1 0.0000
variety:nitrogen  8     0.105   0.013     0.6 0.7756
Residuals        30     0.662   0.022
```

The analysis has detected large differences between Nitrogen levels, and a smaller, but significant difference, between Varieties. The interaction between Nitrogen and Variety is not significant, so the varieties can be assumed to respond similarly to the various levels of Nitrogen. Notice that the Main Plot Error is considerably greater than the Sub Plot Error, which indicates that the use of a split plot design is justified for this experiment, for otherwise the F value for Nitrogen, and its interaction, would have been much lower and differences would not have been detected with the same precision.

3.2 The comparison of treatment means

When significant differences have been detected it is reasonable to ask where these differences occur. There are many methods available for an examination of treatment differences in comparative experiments, but the standard error of the difference between two treatment means is often helpful for this purpose. Its evaluation depends upon the result that if X and Y are independent random variables from the distributions X $\sim$N(μ_1, σ_1^2), Y $\sim$N(μ_2, σ_2^2), and n_1, n_2 independent samples are drawn from these distributions then

$$\bar{X} \; - \; \bar{Y} \sim N\left(\mu_1 \; - \; \mu_2, \; \frac{\sigma_1^2}{n_1} \; + \; \frac{\sigma_2^2}{n_2}\right) \quad .$$

If the variances are equal, then the variance of the difference between two treatment means is $\sigma^2(\frac{1}{n_1} \; + \; \frac{1}{n_2})$ and if the levels of replication are both equal to r the variance of the mean difference between any two treatments is given by $2\frac{\sigma^2}{r}$.
The numbers of levels of M, T and Blocks are m, t and b, the Error variance for main plot is σ_m^2, and for sub plots σ^2. It can be shown that the expected values of the Error Mean Square's for main plots and sub plots are $\sigma^2 \; + \; t\sigma_m^2$ and σ^2.

The error difference between Main Plot Treatments p and q is

$$\bar{\epsilon}_{p.} \ + \ \bar{\bar{\delta}}_{p..} \ - \ \bar{\epsilon}_{q.} \ - \ \bar{\bar{\delta}}_{q..} \ .$$

Since each $\bar{\epsilon}$ is the mean of b values, and each $\bar{\bar{\delta}}$ is the mean of bt values and all errors in this expression are independent, the variance of the difference between Main Plot treatment means is

$$\frac{2}{b}\sigma_m^2 \ + \ \frac{2}{bt}\sigma^2 \ = \ \frac{2}{bt}(\sigma^2 \ + \ t\sigma_m^2) \ .$$

The error difference between Sub Plot treatment means r and s is

$$\bar{\bar{\delta}}_{..r} \ - \ \bar{\bar{\delta}}_{..s}$$

and the Main Plot errors cancel since each Sub Plot treatment occurs in every Main Plot. As each $\bar{\bar{\delta}}$ contains mb independent errors the variance of the difference between Sub Plot treatment means is thus $\frac{2}{mb}\ \sigma^2$.
The interaction is a little more complicated as the error difference is

$$\bar{\epsilon}_{p.} \ + \ \bar{\bar{\delta}}_{p.r} \ - \ \bar{\epsilon}_{q.} \ - \ \bar{\bar{\delta}}_{q.s}$$

Suppose that two treatment combinations have the same main plot treatment, then the ϵ terms cancel as they are exactly the same error term. The variance of the error difference $\delta_{p.q} \ - \ \delta_{r.s}$ is $\frac{2}{b}\sigma^2$. However for different main plots the $\bar{\epsilon}$ terms do not cancel so the variance of the error difference is

$$2\frac{\sigma_m^2}{b} \ + \ 2\frac{\sigma^2}{b} \ = \ \frac{2}{b}\ (\sigma^2 \ + \ \sigma_m^2) \ .$$

Table 3.2: Summary table for variances of treatment differences

Main plot treatments	$\frac{2}{bt}(\sigma^2 \ + \ t\sigma_m^2)$
Sub plot treatments	$\frac{2}{mb}\ \sigma^2$
Combinations(same Main Plot Treatment)	$\frac{2}{b}\sigma^2$
Combinations(different Main Plot Treatment)	$\frac{2}{b}\ (\sigma^2 \ + \ \sigma_m^2)$

In the analysis of variance table the main plot error estimates $\sigma^2 + t\sigma_m^2$ and the sub plot error estimates σ^2. Hence the variance of the difference between T×M means with differing levels of M is

$$\frac{2}{b}(\sigma^2 + \sigma_m^2) \ .$$

It follows therefore that the variance of the T×M treatment difference is

$$\frac{2}{bt} \text{ (Main Plot Error } + \ (t-1)\text{Sub Plot Error)}$$

In the example $b = 4,\ t = 3,\ m = 5$ so $\sigma^2 = 0.022$ and hence $\sigma_m^2 = 0.099$. The Variety means are

V1	V2	V3	V4	V5
5.59	5.24	5.03	5.68	4.81

and so 0.053 is the variance of the difference. It should be remembered that this estimate is not very precise as Varieties were not the main source of interest.

The Nitrogen means are

N1	N2	N3
4.860	5.285	5.665

and the variance of the difference is 0.022. However, since the Nitrogen levels are evenly spaced, it would seem better to examine the Nitrogen effect for linear and quadratic components.

The interaction is not significant, so does not justify close scrutiny. However, the variance of the difference is 0.01103 for Variety×Nitrogen combinations with the same level of Variety, 0.06067 otherwise. An inspection of the two way table of means for Variety and Nitrogen indicates that the response for each factor behaves in a 'parallel' fashion in the presence of the other.

Means table	V1	V2	V3	V4	V5	all V
N1	5.175	4.825	4.050	5.175	4.475	4.860
N2	5.625	5.250	5.050	5.700	4.800	5.285
N3	5.975	5.650	5.400	6.150	5.150	5.665
all N	5.592	5.242	5.033	5.675	4.808	5.270

One point that emerges from this analysis is the comparison of the error variances. For comparisons of Sub Plot treatments the error variance is σ^2, but if the mt treatments had been arranged in b randomised blocks a much greater error variance could have been expected since the Main Plot error is 14.5 times as great as the Sub Plot error. Some of the comparisons that can be made e.g. of Main Plot treatment levels at fixed levels of the Sub Plot treatment, or across different Main Plot and Sub Plot treatment levels, have major inferential problems. These are due to the estimates of variance involving two different Residual Mean Squares with different $d.f.$ and different expectations.

3.3 The mixed model analysis

See [12].

For a vector of observations from block i and mainplot i,j, we write the model as

$$\begin{aligned} \mathbf{y}_{ij} &= \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{1,ij}\mathbf{b}_i + \mathbf{Z}_{2,ij}\mathbf{d}_{ij} + \boldsymbol{\epsilon}_{ij} \\ \mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\psi}_1)\ , \mathbf{d}_{ij} &\sim N(\mathbf{0}, \boldsymbol{\psi}_2)\ \ \boldsymbol{\epsilon}_{ij} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned} \tag{3.1}$$

Furthermore, the random effects are uncorrelated.

The design matrix for blocks is $\mathbf{Z}_1$, $\mathbf{Z}_2$ is the design matrix for the block by main treatment interaction, $\boldsymbol{\psi}_2$ is diag(σ_m^2),.

In the lme() function, we include a random term for blocks and main plots (variety in this case) within blocks,

```
Wheat <- expand.grid(nitro=1:3,Variety=paste("V",1:5),block=1:4)

Wheat$block <- factor(Wheat$block)
Wheat$Variety <- factor(Wheat$Variety)
Wheat$nitro <- factor(Wheat$nitro)


Wheat$yield <- c(5.5,6.1,6.4,4.6,5.3,5.5,4.9,5.5,5.8,5.0,5.7,6.0,4.7,5.0,5.4,
           5.0,5.5,5.8,4.7,5.0,5.4,4.9,5.4,5.5,5.7,6.2,6.5,4.2,4.6,5.0,
           5.0,5.4,5.9,5.0,5.1,5.8,4.8,4.7,5.5,5.3,5.8,6.7,4.6,4.8,5.0,
           5.2,5.5,5.8,5.0,5.6,5.9,4.0,4.6,4.8,4.7,5.1,5.4,4.4,4.8,5.2)
library(nlme)
Wheat.model <- lme(yield ~ nitro*Variety,random= ~1 | block/Variety,data=Wheat)

print(VarCorr(Wheat.model))

            Variance     StdDev
block =     pdLogChol(1)
(Intercept) 0.001184     0.03441
Variety =   pdLogChol(1)
(Intercept) 0.099120     0.31483
Residual    0.022054     0.14850
```

If we equate the observed mean squares to their expected values using the results in Table 3.1,

$$\begin{aligned} \sigma_m^2 + 3 \times \sigma^2 &= 0.320 \\ \sigma^2 &= 0.022 \end{aligned}$$

Hence the estimates of variances are:-

$$\begin{aligned} \hat{\sigma}^2 &= 0.022 \\ \hat{\sigma}_m^2 &= \frac{(0.32 - 0.022)}{3} = 0.1 \end{aligned}$$

and these estimates (Moment estimates) agree with the mixed model estimates.

Chapter 4

Mixed Model Theory

The linear mixed model has the general form,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \mathbf{Zu} + \epsilon & (4.1)\\ &= \left[\begin{array}{c|c} \mathbf{X} & \mathbf{Z} \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \mathbf{u} \end{array}\right] + \epsilon & (4.2) \end{aligned}$$

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{D}), \qquad \epsilon \sim N(\mathbf{0}, \mathbf{R}) \quad , E(\mathbf{u}\boldsymbol{\epsilon}) = 0$$

4.1 Residual Maximum Likelihood

Temporarily consider model (4.1) as

$$y = X\beta + \varepsilon \quad , \varepsilon \sim N(0, \sigma^2 V) \tag{4.3}$$

where

$$\begin{aligned} \varepsilon &= \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \ldots + \boldsymbol{\epsilon} \\ \mathbf{V} &= \left(I + \sum_{i=1}^{n_u} \gamma_i \mathbf{Z}_i \mathbf{Z}_i^T\right) \quad \gamma_i = \sigma_i^2/\sigma_e^2 \quad , E(u_i u_j) = 0, \quad E(u_i, \epsilon) = 0 \end{aligned}$$

The purpose of rewriting the model this way is to indicate that the variance components are from the space of **error contrasts**.

For known Z_i, the variance matrix contains unknown parameters γ_i and the dispersion parameter σ^2. In a model with only 1 random component, the measurement error, it is not necessary to model the random components because they are residuals. With more than 1 random component, we need to model sources of the variance.

In mixed models we estimate variance components for

(i) efficiency of fixed effects,

(ii) reliable estimates of fixed effects,

(iii) of interest in themselves, not nuisance parameters.

In a model containing only random effects, the variance components can be estimated by equating observed mean squares to their expected values. In a balanced mixed effects model, the old way was to estimate fixed effects, subtract from the data and apply moment estimation to the residuals, (Henderson's methods). In cases where there is a lack of balance, matrix multiplication becomes exorbitant so we turn to likelihood methods.

The normal likelihood function is

$$L(\beta, \gamma, \sigma|\mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}}|V|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T V^{-1}(y - X\beta)\right\}$$

and its logarithm is

$$\ell(\beta, \gamma, \sigma|\mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|V| - \frac{1}{2\sigma^2}(y - X\beta)^T V^{-1}(y - X\beta) \ . \tag{4.4}$$

Then,

$$\frac{\partial \ell}{\partial \beta} = 0 \quad \Rightarrow \hat{\beta} \quad = (X^T V^{-1} X)^- X^T V^{-1} y \tag{4.5}$$

$$\frac{\partial \ell}{\partial \gamma_i} = 0 \quad \Rightarrow \hat{\gamma}_i \quad = \tag{4.6}$$

$$\frac{\partial \ell}{\partial \sigma^2} = 0 \quad \Rightarrow \hat{\sigma}^2 \quad = n^{-1}(Y - X\beta)^T \sigma^{-1}(y - X\beta) \tag{4.7}$$

The equations are solved by the iterative Newton-Raphson algorithm,

$$\begin{aligned} \begin{bmatrix} \beta \\ \gamma_1 \\ \vdots \\ \sigma^2 \end{bmatrix}_j &= \begin{bmatrix} \beta \\ \gamma_1 \\ \vdots \\ \sigma^2 \end{bmatrix}_{j-1} - \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta^2} & \cdots & & \\ \frac{\partial^2 \ell}{\partial \beta \partial \gamma_1} & \frac{\partial^2 \ell}{\partial \gamma_1^2} & \cdots & \\ \vdots & \vdots & \ddots & \\ & & & \frac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{bmatrix}_{j-1}^{-1} \begin{bmatrix} \frac{\partial \ell}{\partial \beta} \\ \frac{\partial \ell}{\partial \gamma_1} \\ \vdots \\ \frac{\partial \ell}{\partial \sigma^2} \end{bmatrix}_{j-1} \\ \boldsymbol{\theta}_j &= \boldsymbol{\theta}_{j-1} - \mathcal{I}_{j-1} \times \boldsymbol{\Delta}_{j-1} \end{aligned} \tag{4.8}$$

The matrix of second derivatives is called the *Hessian* and the negative inverse is the *observed information matrix*, $\mathcal{I}$. The expectation of the negative of the inverse Hessian is the *Fisher Information matrix*.

In (4.5) and (4.7), we see that an estimate of β is required to get an estimate of σ^2 which is a "double dipping" of the information contained in the data. Residual Maximum Likelihood arose from the observation that estimates of variance components in mixed model estimation were often biassed downwards whereas they were not for the variance component model (ie all random effects).

Patterson and Thompson [6] subdivided the likelihood into 2 spaces,

(i) the space spanned by treatment contrasts, and

(ii) the space spanned by error contrasts, orthogonal to the treatment contrasts.

The treatment given below is well explained in Verbyla (1990) [13].
For $L = (L_1, L_2)^T$ such that $L_1^T X = I_p$ and L_2 such that $L_2^T X = \mathbf{0}_{(n-p)}$,

$$L^T \mathbf{y} = \begin{bmatrix} L_1^T \\ L_2^T \end{bmatrix} \times \mathbf{y} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \end{bmatrix}$$

then,

$$\begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{B}_p \\ \mathbf{0}_{(n-p)} \end{bmatrix}, \sigma^2 \begin{bmatrix} L_1 V L_1^T & L_1 V L_2^T \\ L_2 V L_1^T & L_2 V L_2^T \end{bmatrix} \right)$$

One choice for L_1 (of rank p) is $(X^T X)^{-1} X^T$.
The likelihood can be written as the sum of orthogonal components,

$$\ell(\beta, \gamma, \sigma | \mathbf{y}) = \underbrace{\ell_1(\beta | \mathbf{y}_1)}_{\text{fixed}} + \underbrace{\ell_2(\gamma, \sigma | \mathbf{y}_2)}_{\text{set of error contrasts}} .$$

Variance parameter estimates are estimated by maximizing ℓ_2 while estimation of β is through ℓ_1.

The forms of $-2\log(\ell_2)$ and $-2\log(\ell)$ are given below in (4.9) and (4.10),

$$-2\log(\ell_2) = (n-p)\log(2\pi) \quad - \log|X^T X| \quad + \log|V| + \log|X^T V^{-1} X| + (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \tag{4.9}$$

$$-2\log(\ell) = (n-p)\log(2\pi) \qquad\qquad + \log|V| + \log|X^T V^{-1} X| + (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \tag{4.10}$$

The difference between (4.9) and (4.10) is the term $\log|X^T X|$ which penalises the likelihood of the nuisance parameters whose information is in y_2 for the degrees of freedom used in calculating the interest parameters, $\hat{\beta}$.

4.2 Properties of REML

(i) For every set of $(N-p)$ linear error contrasts, the REML log-likelihood is the same.

(ii) No matter what $(N-p)$ linearly independent error contrasts are used, maximising their likelihood always leads to the same equations for estimating variance components γ_i and σ^2.

(iii) Optimal property. The derivation which shows ℓ_2 to be a marginal likelihood for $Y_2 = L_2^T y$ also shows that ℓ_2 is the conditional likelihood of $y|\hat{\beta}$.

(a) For given (γ_i, σ^2), the estimator $\hat{\beta}$ is sufficient for β, so it is this conditional likelihood which is used to find the most powerful similar tests of hypotheses concerning (γ_i, σ^2).

(b) estimators $\hat{\beta}$,s^2 are jointly sufficient for β,σ^2 for any given γ so that most powerful similar tests for γ alone may be constructed using the conditional distribution of $(y|\hat{\beta}, s^2)$. Now $s^2/\sigma^2 \sim \chi^2_{n-p}$ so that the density function of s^2,

$$f_1(s^2) = \frac{s^{n-p-2} \exp\left\{-\frac{s^2}{2\sigma^4}\right\}}{\sigma^{n-p} 2^{\frac{n-p}{2}} \Gamma\left(\frac{n-p}{2}\right)}$$

and the distribution of $(y|\hat{\beta}, s^2)$,

$$f_2(y|\hat{\beta}, s^2) \propto \exp\left\{\log|V| + \log|X^T V^{-1} X| + (n-p-2)\log s^2\right\}$$

does not involve β, σ^2. At $\gamma = \gamma_0$, $V = V(\gamma_0)$,

$$\begin{aligned} \hat{\beta} &= (X^T V^{-1} X)^- X^T V^{-1} y \\ s^2 &= (y - X\hat{\beta})^T V^{-1} (y - X\hat{\beta}) \end{aligned}$$

A most powerful similar test of

$$\begin{aligned} H_0 : \gamma &= \gamma_0 \qquad \text{versus} \\ H_1 : \gamma &= \gamma_1 \end{aligned}$$

is $\ell_3(\gamma_1) - \ell_3(\gamma_0) \geq K$ where K is chosen to give some significant level.

(iv) Unbiassed estimating equations.

$$\ell = -\frac{1}{2}\left[n\log(2\pi\sigma^2) + \log|V| + \sigma^{-2}(y - X\beta)^T V^{-1}(y - X\beta)\right]$$

where $V = V(\theta)$. The MLE's are derived by equating derivatives to zero,

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \sigma^{-2} X^T V^{-1}(y - X\beta) \\ \frac{\partial \ell}{\partial \sigma^2} &= -(n/2\sigma^2) + (1/2\sigma^4)(y - X\beta)^T V^{-1}(y - X\beta) \\ \frac{\partial \ell}{\partial \gamma_i} &= -\frac{1}{2}\left\{\frac{\partial \log|V|}{\partial \gamma_i} + \sigma^{-2}(y - X\beta)^T \left[\frac{\partial V^{-1}}{\partial \gamma_i}\right](y - X\beta)\right\} \end{aligned}$$

The expectations of each of these expresions is zero. Moreover $E\left(\frac{\partial \ell}{\partial \beta}\right) = 0$ when $\hat{\sigma}^2$, $\hat{\gamma}$ is substituted for σ^2, γ. Thus $\frac{\partial \ell}{\partial \beta}$ remains an unbiassed estimating equation giving

$$\hat{\beta} = (X^T V^{-1} X)^- X^T V^{-1} y \ .$$

However if $\hat{\beta}$ is substituted for β in $\frac{\partial \ell}{\partial \sigma^2}$ and $\frac{\partial \ell}{\partial \gamma_i}$, the expectation is no longer zero,

$$E\left\{(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta})\right\} = (n-p)\sigma^2 \quad .$$

Therefore,

$$E\left(\frac{\partial \ell}{\partial \sigma^2}\right) = -\frac{n}{2\sigma^2} + \frac{(n-p)\sigma^2}{2\sigma^4} = -\frac{n}{2\sigma^2} + \frac{(n-p)}{2\sigma^2} \ .$$

Hence,

$$\frac{\partial \ell}{\partial \sigma^2} - E\left(\frac{\partial \ell}{\partial \sigma^2}\right) = -\frac{(n-p)}{2\sigma^2} + \frac{(y - X\hat{\beta})^T V^{-1}(y - X\hat{\beta})}{2\sigma^4}$$

becomes the unbiassed estimating equation for true σ^2 used in V^{-1}. For estimating γ_i, first note that

$$E\left\{(y - X\hat{\beta})(y - X\hat{\beta})^T\right\} = \sigma^2\left[V - X(X^T V^{-1} X)^{-1} X^T\right] \ .$$

The derivative for γ_i has expectation given by,

$$\begin{aligned}
-2E\left(\frac{\partial \ell}{\partial \gamma_i}\right) &= \frac{\partial \log|V|}{\partial \gamma_i} + \sigma^{-2} \times E\left\{tr\left[\frac{\partial}{\partial \gamma_i}(V^{-1}(y - X\hat{\beta})(y - X\hat{\beta})^T\right]\right\} \\
&= \frac{\partial \log|V|}{\partial \gamma_i} + tr\left\{\frac{\partial V^{-1}}{\partial \gamma_i}\left[V - X(X^T V^{-1} X)^{-1} X^T\right]\right\} \\
&= -tr\left[X^T\left(\frac{\partial V^{-1}}{\partial \gamma_i}\right) X (X^T V^{-1} X)^{-1}\right] \\
&= \frac{\partial\left(\log|X^T V^{-1} X|\right)}{\partial \gamma_i} \ .
\end{aligned}$$

The unbiassed estimating equation after β is replaced by $\hat{\beta}$ in $\frac{\partial \ell}{\partial \gamma_i}$ is

$$\frac{\partial}{\partial \gamma_i}\log|V| + \frac{\partial}{\partial \gamma_i}\log|X^T V^{-1} X| + \sigma^{-2}(y - X\hat{\beta})^T \frac{\partial V^{-1}}{\partial \gamma_i}(y - X\hat{\beta}) = 0$$

which is the same as in previous expressions of REML.

The construction of estimating equations via

$$\frac{\partial \ell}{\partial \theta} - E\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

is known as profile likelihood estimation [1].

Write

$$\frac{\partial \ell^\star}{\partial \theta} = \frac{\partial \ell}{\partial \theta} - E\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

$$\text{ie. } \ell^\star = \ell - \psi$$

$$\text{where } \psi = \int E\left(\frac{\partial \ell}{\partial \theta}\right) d\theta \ .$$

Hence ψ can be considered an adjustment term for the log-likelihood.

For normal distributions, the adjustment is

$$\psi = -\frac{1}{2}\log\{2\pi \text{var}(\hat{\beta})\} = \frac{1}{2}\log|2\pi\mathcal{I}_{\hat{\beta}}|$$

where $\mathcal{I}_{\hat{\beta}}|$ is the information matrix for $\hat{\beta}$ or $X^TV^{-1}X$.

4.3 Orthogonality of parameters

Our aim is to estimate the conditional distribution of the interest parameters, given the MLE for nuisance parameters.

In (4.2), the parameter set is partitioned into 2 vectors of lengths p_1, p_2. In the mixed model we have stipulated that random effects are uncorrelated. Cox and Reid (1987) define the orthogonality property for $\theta = (\psi, \lambda)$, of lengths (p_1, p_2), by

$$i_{\psi,\lambda} = \frac{1}{n}E\left(\frac{\partial\ell}{\partial\psi}\frac{\partial\ell}{\partial\lambda}\right) = \frac{1}{n}E\left(-\frac{\partial^2\ell}{\partial\psi\partial\lambda}\right) = 0 \tag{4.11}$$

where $i_{\psi,\lambda}$ is an element of the Information matrix. Orthogonality between interest and nuisance effects is a key condition of the REML method, even if it is local rather than global orthogonality. The element i is assumed to be $\mathcal{O}(1)$ as $n \to \infty$.[1]

If (4.11) holds at only 1 parameter value θ^0, it is locally orthogonal at θ^0. Local orthogonality can always be achieved in a Hilbert space but global orthogonality is possible only in special cases.

For $\theta = (\psi, \lambda)$,

(a) the MLE's, $\hat{\psi}$ and $\hat{\lambda}$ are asymptotically independent,

(b) se$(\hat{\psi})$ is the same whether λ is known or not,

(c) the estimation of $(\hat{\psi}, \hat{\lambda})$ is simpler,

(d) MLE of ψ when λ is known, $\hat{\psi}_\lambda$, varies only slowly with λ.

To study the last point, expand the joint likelihood near $(\hat{\psi}, \hat{\lambda})$,

$$\begin{aligned}
\ell(\psi,\lambda) &= \ell(\hat{\psi},\hat{\lambda}) + \underbrace{\left[\begin{array}{cc}(\psi-\hat{\psi}) & (\lambda-\hat{\lambda})\end{array}\right]\left[\begin{array}{c}\frac{\partial\ell}{\partial\psi}\\ \frac{\partial\ell}{\partial\lambda}\end{array}\right]_{\hat{\psi},\hat{\lambda}}}_{=0} + \\
&\quad \frac{1}{2}\left[\begin{array}{cc}(\psi-\hat{\psi}) & (\lambda-\hat{\lambda})\end{array}\right]\left[\begin{array}{cc}\frac{\partial^2\ell}{\partial\psi^2} & \frac{\partial^2\ell}{\partial\psi\partial\lambda}\\ \frac{\partial^2\ell}{\partial\psi\partial\lambda} & \frac{\partial^2\ell}{\partial\lambda^2}\end{array}\right]_{\hat{\psi},\hat{\lambda}}\left[\begin{array}{c}(\psi-\hat{\psi})\\ (\lambda-\hat{\lambda})\end{array}\right] + \\
&\quad \mathcal{O}_p\left(\left\|\left[\begin{array}{cc}(\psi-\hat{\psi}) & (\lambda-\hat{\lambda})\end{array}\right]\right\|^3\right) \\
&\approx \ell(\hat{\psi},\hat{\lambda}) + \frac{1}{2}\left[-n\hat{j}_{\psi,\psi}(\psi-\hat{\psi})^2 - 2n\hat{j}_{\psi\lambda}(\psi-\hat{\psi})(\lambda-\hat{\lambda}) - n\hat{j}_{\lambda\lambda}(\lambda-\hat{\lambda})^2\right]
\end{aligned} \tag{4.12}$$

[1] $\mathcal{O}(1)$ is explained in the appendix.

where $\hat{j}_{\psi\psi} = i_{\psi\psi} + Z_{\psi\psi}/\sqrt{n}$, $E(Z_{\psi\psi}) = 0$ and $Z_{\psi\psi} = \mathcal{O}_p(1)$.

Rewriting (4.12) in terms of i and Z and differentiating wrt ψ, $\hat{\psi}_\lambda$ satisfies the relation

$$ni_{\psi\psi}(\hat{\psi}_\lambda - \hat{\psi}) + \sqrt{n}Z_{\psi\psi}(\hat{\psi}_\lambda - \hat{\psi}) + \frac{1}{2}(\hat{\psi}_\lambda - \hat{\psi})^2\sqrt{n}\frac{\partial Z_{\psi\psi}}{\partial\psi} + \frac{1}{2}(\hat{\psi}_\lambda - \hat{\psi})^2 n\frac{\partial i_{\psi\psi}}{\partial\psi} + \ldots = 0$$

Because of the following results ,

$$\begin{array}{rcl} \frac{\partial Z_{\psi\psi}}{\partial\psi} & = & \mathcal{O}_p(1) \\ \frac{\partial i_{\psi\psi}}{\partial\psi} & = & \mathcal{O}(1) \\ \hat{\psi}_\lambda - \hat{\psi} & = & \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right) \\ \hat{\lambda}_\psi - \hat{\lambda} & = & \mathcal{O}_p\left(\frac{1}{n}\right) \end{array},$$

then iff $i_{\psi\lambda} = 0$ (ie orthogonality), the first term is $\mathcal{O}_p(\sqrt{n})$ and the remaining terms are $\mathcal{O}(1)$.

The idea that $\hat{\psi}_\lambda$ is similar for different λ is illustrated in Figure 4.1 where $\psi = \nu$ and $\lambda = \sigma^2$.

If $\hat{\psi}_\lambda = \hat{\psi}$ for all λ, then ψ and λ are orthogonal parameters. Models from the exponential family contain ψ as part of the canonical parameter and λ as the complementary part of the expectation, eg.

- Normal μ, σ^2
- Gamma ν, ϕ (shape and scale parameters)

Typically ψ is orthogonal to $\lambda_1, \ldots, \lambda_{p_2}$ where ψ is interest and the λ's are nuisance but they can be of interest. This is the basis of REML - finding a representation of the nuisance effects orthogonal to the interest effects.

4.4 Mixed Model Equations (MME)

From (4.1), we have

$$\begin{array}{rrll} \text{var}(\mathbf{y}) & = \mathbf{V} & = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R} & (4.13) \\ \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\hat{\beta} & = & \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} & (4.14) \end{array}$$

The estimating equations for the mixed model are

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{B}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (4.15)$$

Because the equations are solved iteratively, $\mathbf{D}$ may become singular and a preferable alternative to (4.15) is

$$\begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + I \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{B}} \\ \tilde{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (4.16)$$

Figure 4.1: Estimation of interest, given nuisance parameters and orthogonality

177

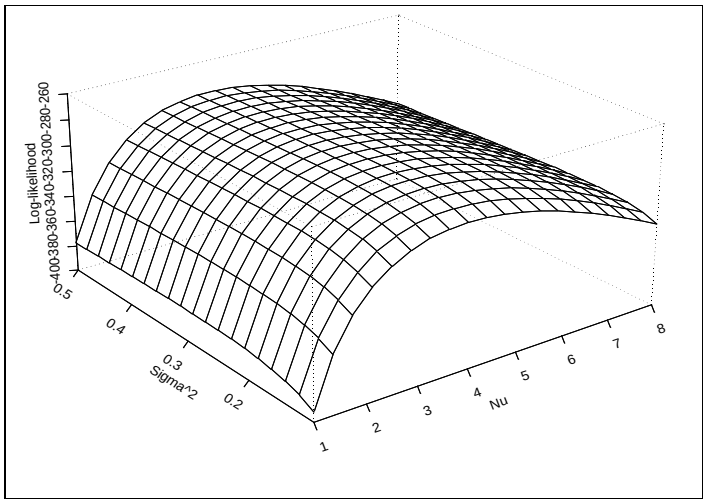


Figure 7.3: *Perspective plot of the log-likelihood against the dispersions ν and σ^2. The parameter values are $\nu = 4$ and $\sigma^2 = 0.25$. The plot shows that the peak with respect to σ^2 is not as noticeable as its counterpart ν*

where $\mathbf{D}\tilde{\nu} = \tilde{u}$.

Taking the matrix results on faith,

$$\begin{bmatrix} \hat{\mathbf{B}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} (XV^{-1}X)^{-}X^{T}V^{-1}\mathbf{y} \\ DZ^{T}V^{-1}(\mathbf{y} - X\hat{\beta}) \end{bmatrix} \tag{4.17}$$

The algorithm is

1. Derive variance components of $\mathbf{V}$ by REML.
2. Given the variance parameters, calculate $\hat{\mathbf{B}}$ by Generalised Least Squares (GLS).
3. Given variance components and $\hat{\mathbf{B}}$, calculate $\tilde{\mathbf{u}}$.

Some points of note:-

(i) $\tilde{u}$ is referred to as the Best Linear Unbiased Predictor (BLUP).

Although $\hat{\mathbf{B}}$ is not invariant to the choice of $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$, the occurrence of $\hat{\mathbf{B}}$ in $\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{B}})$ is such that $\tilde{u}$ is invariant to $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ and $\tilde{u}$ is the same for all solutions of $\hat{\mathbf{B}}$.

(ii) Under normality assumptions, $\tilde{u}$ is exactly the same as $E(u|y)$, save for $\hat{\beta}$ in place of β. Since under normality $\hat{\beta}$ is the m.l.e. of β when $\mathbf{D}$ and $\mathbf{V}$ are known, we can refer to

$$\tilde{\mathbf{u}} = \widehat{E(\mathbf{u}|\mathbf{y})} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{B}})$$

as the ML estimator of the conditional mean $E(u|y)$.

Chapter 5

Variance Models

Systematic effects are modelled by curves, means etc. and random efects are modelled by density functions. At the heart of the multivariate density function is the covariance matrix. In the same way as we seek to generalise the systematic effect through models, we also find "smooth" models for densities to conserve degrees of freedom yet retain the important information.

Before launching into a mixed-model analysis, recall that there must be repeated measures (in time, space or measurement) from a sampling unit to have a mixed model with random subject effects. That means we have to mentally organise the data according to its source, ie. subject, repeated measures, treatment etc.

5.1 Variance Components

Write the mixed model with separate the mean and covariance structures,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} \tag{5.1}$$

and

$$\mathbf{E} \sim N\left(\mathbf{0}, \mathbf{V}(\alpha)\right) \tag{5.2}$$

We assume that $\mathbf{E}$ can be additively decomposed into

1. random effects,
2. serially correlated variation and
3. measurement error,

$$\mathbf{E}_{i,j} = \mathbf{Z}_i\mathbf{u}_i + \mathbf{W}_i(t_{ij}) + \epsilon_{i,j} \ , \tag{5.3}$$

and

$$\text{var}(\mathbf{E}_{i,j}) = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \sigma_r^2\mathbf{R} + \sigma^2\mathbf{I} \ . \tag{5.4}$$

We decompose the systematic part of the model into components guided by the experiment design and exploratory data analysis. Once we recognise the structure of the random part, we can explore for likely components.

5.1.1 Pure Serial Correlation

Assume that there are repeated measurements over time from a single subject,

$$y_1, y_2, \ldots, y_n \ ,$$

so that equation (5.3) reduces to

$$\mathbf{E}_{i,j} = \mathbf{W}_i(t_{ij})$$

and (5.4) is

$$\text{var}(\mathbf{E}_{i,j}) = \sigma^2\mathbf{R}$$

The sample covariance amongst the $E_{i,j}$ is

$$\begin{bmatrix} \sigma_{11} & & & & \\ \sigma_{21} & \sigma_{22} & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \sigma_{n1} & & & & \sigma_{nn} \end{bmatrix} = \sigma^2 \begin{bmatrix} \rho_{11} & & & & \\ \rho_{21} & \rho_{22} & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ \rho_{n1} & & & & \rho_{nn} \end{bmatrix}$$

If we use an outo-correlation function $\rho(u)$ to model the correlation of observations that are u apart, eg. $\rho(u) = \exp(-\phi u)$, the covariance matrix is modelled in terms of the parameter ϕ.

$$\mathbf{V} = \sigma^2 \begin{bmatrix} \rho(0) & & & & \\ \rho(1) & \rho(0) & & & \\ \rho(2) & \rho(1) & \ddots & & \\ \vdots & & & \ddots & \\ \rho(n-1) & & & \rho(1) & \rho(0) \end{bmatrix}$$

5.1.2 Random effects plus measurement error

The components from equation (5.3) are

$$\mathbf{E}_{i,j} = \mathbf{Z}_i\mathbf{u}_i + \epsilon_{i,j} \ .$$

Random Intercepts

If the random effects are scalar intercepts,

$$\text{var}(\mathbf{E}) = \sigma^2 \left(\frac{\sigma_u^2}{\sigma^2} J + I \right) \ .$$

The variance of each E_j is $\sigma^2 + \sigma_u^2$ and the correlation between any 2 measurements from the same unit is $\rho = \sigma_u^2/(\sigma^2 + \sigma_u^2)$.

Random Slopes

If

$$\mathbf{Y} = \mathbf{X}\beta \quad + \overbrace{\begin{bmatrix} 1 & t_1 & \\ 1 & t_2 & \\ \vdots & \vdots & \\ 1 & t_K & \\ & 1 & t_1 \\ & 1 & t_2 \\ & \vdots & \vdots \\ & 1 & t_K \end{bmatrix}}^{\mathbf{Z}} \begin{bmatrix} \mathbf{u}_{11} \\ \mathbf{u}_{12} \\ \mathbf{u}_{21} \\ \mathbf{u}_{22} \end{bmatrix} + \epsilon$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are Normal random vectors with variances σ_1^2 and σ_2^2, representing intercepts and slopes respectively,

$$\text{var}(E_j) = \sigma_1^2 + t_j^2\sigma_2^2 + \sigma^2 \ .$$

For $j \neq k$,

$$\text{Cov}(E_j, E_k) = \sigma_1^2 + t_j t_k \sigma_2^2$$

where we see that covariance increases as t increases.

Correlation may be represented by random effects

Diggle, Liang and Zeger (1994), page 88 state

> Whilst serial correlation would appear to be a natural feature of any longitudinal model, in specific applications its effect may be dominated by the combination of random effects and measurement error.

5.1.3 Random effects, Serial Correlation and Measurement Error

If in the above example the time variable was not included in the $\mathbf{Z}$ matrix, the variance model would have components due to random intercepts, measurement error and as likely, there would be serial correlation.

The variance matrix (5.4) becomes

$$\text{var}(\mathbf{E}) = \sigma^2 \left(I + \frac{\sigma_u^2}{\sigma^2} J + \frac{\sigma_r^2}{\sigma^2} R \right) \ , \tag{5.5}$$

where

- I is the identity matrix,
- J is a matrix of 1's or $\mathbf{11}'$,
- R is a correlation matrix.

The information on σ_u^2 comes from replication of treatment units, σ_r^2 is derived from amongst times within units and σ^2 is residual.

5.1.4 The variogram

The diagnostic to show the relative importance of each of the random components is the variogram (see [2]), defined as

$$\gamma(v) = \frac{1}{2}E\left[\{Y(t) - Y(t-v)\}^2\right],\ v \geq 0$$

In a first pass at the data, a saturated model is fitted for the fixed effects, saving the residuals. The sample variogram is a plot of the observed half squared differences between pairs of residuals (v_{ijk}) and plotted versus the corresponding time (or space) differences (δ_{ijk}),

$$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2, \qquad \delta_{ijk} = t_{ij} - t_{ik} \quad .$$

A smooth line regressed through the points (δ, v).

The information contained in the v_{ijk} would have the contributions from

- Measurement error,
- Components due to random effects,
- Serial correlation.

Figure 5.1 is an idealised variogram which shows how the components affect the shape of the curve. From the sample variogram of the residuals, we can gauge what variance components should be modelled and when the variance model is satisfactory, the variogram should be flat.

Figure 5.1: Variogram

5.2 Matrix results

5.2.1 "Tearing"

Solution for $\hat{\mathbf{B}}$ requires inverting $\mathbf{V}$ but as $\mathbf{D}$ is often block diagonal and $\mathbf{R}$ either diagonal or patterned (eg banded), we use the result that

$$\begin{aligned} \mathbf{V_{n,n}}^{-1} &= (\mathbf{Z}_{n,q}\mathbf{D}_{q,q}\mathbf{Z}_{q,n}^T + \mathbf{R})_{n,n}^{-1} \\ &= \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{D}^{-1} + \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{R}^{-1} \end{aligned} \tag{5.6}$$

Now if $\mathbf{R}$ has a "simple" inverse and $q << n$, the patterned form is computationally less intensive than the original form.

5.2.2 Kronecker Product

Let $A = [a_{ij}]$ be an $m \times n$ matrix, $B = [b_{kl}]$ be an $p \times q$ matrix,

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}$$

In this product matrix the term $a_{ij}b_{kl}$ occurs in row $(i-1)p+k$ and column $(j-1)q+1$ but it is usually easier to refer to such a term as the $i, j; k, l$ term since it is the k, l term of block i, j.

5.2.3 The vec operator

The *vec* of an $m \times n$ matrix A, denoted by vec(A), is an mn vector formed by packing the columns of the matrix A one below the other.

If A is $m \times n$, B is $n \times p$, C is $p \times q$, then

$$\text{vec}(AB) = (I \otimes A)\text{vec}B = (B' \otimes I_m)\text{vec}A \tag{5.7}$$

$$\text{vec}(ABC) = (C' \otimes A)\text{vec}B \tag{5.8}$$

5.2.4 Cholesky Decomposition

For any $n \times n$ symmetric positive definite matrix A, there exists a unique diagonal matrix triangular T with positive diagonal elements such that

$$A = TT' \tag{5.9}$$

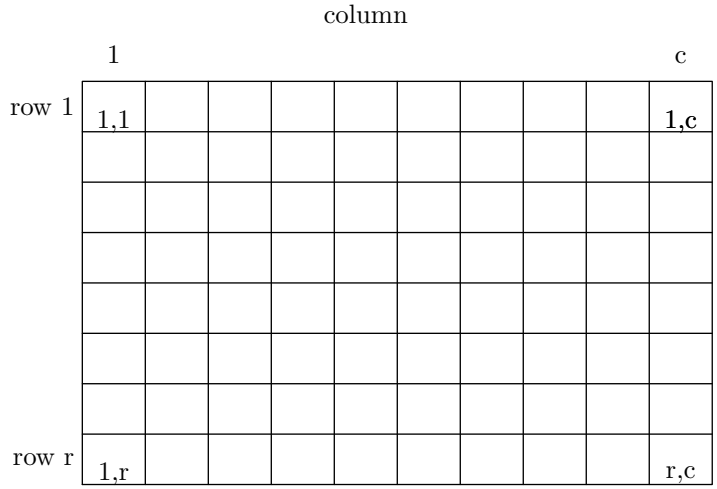
Moreover, taking U to be the unique upper triangular matrix and $D = \{d_i\}$ to be the unique diagonal matrix such that $A = U'DU$,

$$T = D^{\frac{1}{2}}U \tag{5.10}$$

where $D^{\frac{1}{2}} = \text{diag}(\sqrt{d_1}, \sqrt{d_2}, \dots \sqrt{d_n})$.

5.3 Nested Random Effects

Consider a field trial where spatial variation can be anticipated and where yields from adjacent plots may be more alike (ie. higher correlation) than when the plots are more widely separated.



One model would be to regard rows and columns as random effects,

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1c} \\ \\ \\ \\ \\ \\ Y_{r1} \\ \vdots \\ Y_{rc} \end{bmatrix} = \mathbf{X}\beta + \begin{bmatrix} 1 & & \\ \vdots & & \\ 1 & & \\ & 1 & \\ & \vdots & \\ & 1 & \\ & & \\ & & \\ & & 1 \\ & & \vdots \\ & & 1 \end{bmatrix} \times \begin{pmatrix} Ur_1 \\ \vdots \\ Ur_r \end{pmatrix} + \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ 1 & & \\ & \ddots & \\ & & 1 \\ & & \\ & & \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \times \begin{pmatrix} Uc_1 \\ \vdots \\ Uc_c \end{pmatrix} + \epsilon$$

Rather than estimating individual row and column effects (which are nuisance), we could detrend the spatial variation amongst residuals across the field with a model of random effects. This has the effect of mathematically adjusting the yields from individual plots to what they would be on the average plot, ie. subtract the plot effect.

If the correlation amongst rows is $\rho_r(|i-j|)$ and columns is $\rho_c(|m-n|)$, the correlation between 2 plots which are d_r rows apart and d_c columns apart may be reasonably represented by $\rho_r(d_r) \times \rho_c(d_c)$; this is called a separable process. The $rc \times rc$ correlation matrix is

$$\Psi_{rc,rc} = R_{r,r} \otimes C_{c,c} \ .$$

So options are

(i) Include row and column random effects

(ii) Detrend by modelling the serial correlation across the plots.

Doing both (i) and (ii) would likely be overfitting.

5.4 Patterned Matrices in R

Modelling the variances leads to patterned variance-covariance matrices. In R , the `nlme` library provides several classes of positive definite matrices that are used to specify patterned variance-covariance matrices for random effects. They are

`pdBlocked`	blocked diagonal
`pdCompSymm`	split-plot
`pdDiag`	diagonal
`pdIdent`	multiple of Identity
`pdSymm` (default)	general +'ve definite matrix

5.4.1 Split-plot experiment on Oats - alternative way 1

Let indices i, j, k indicate Block i, Variety j, Nitrogen k.

For the ith block,

$$\overbrace{\begin{bmatrix} Y_{i11} \\ Y_{i12} \\ Y_{i13} \\ Y_{i14} \\ Y_{i21} \\ Y_{i22} \\ Y_{i23} \\ Y_{i24} \\ Y_{i31} \\ Y_{i32} \\ Y_{i33} \\ Y_{i24} \end{bmatrix}}^{\mathbf{y_i}} = \overbrace{\begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & N_3 & \\ & & & N_4 \\ N_1 & & & \\ & N_2 & & \\ & & N_3 & \\ & & & N_4 \\ N_1 & & & \\ & N_2 & & \\ & & N_3 & \\ & & & N_4 \end{bmatrix}}^{\mathbf{X}} \times \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \overbrace{\begin{bmatrix} 1 & \\ 1 & \\ 1 & \\ 1 & \\ & 1 \\ & 1 \\ & 1 \\ & 1 \\ & 1 \\ & 1 \\ & 1 \\ & 1 \end{bmatrix}}^{\mathbf{Z}} \times \begin{bmatrix} b_i + b_{i,1} \\ b_i + b_{i,2} \\ b_i + b_{i,3} \end{bmatrix} + \epsilon$$

Note,

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & N_3 & \\ & & & N_4 \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} 1 & \\ & 1 \\ & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

Define distributions,

$$\mathbf{b_i}^* \sim N(0, \psi_\mathbf{i}^*) \ \ , \epsilon_\mathbf{i} \sim N(0, \sigma^2\mathbf{I}) \ \ , \mathrm{cov}(\mathbf{b_i}^*, \epsilon_\mathbf{i}) = 0 \ ,$$

where

$$\psi_\mathbf{i}^* = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{bmatrix} .$$

The $\psi_\mathbf{i}^*$ matrix is compound symmetric.

The R code for this model and the output are listed below.

```
library(nlme)
data(Oats)
mm1.Oats <- lme(yield ~ nitro,data=Oats,
                  random=list(Block=pdCompSymm(~Variety-1) ) )

print(summary(mm1.Oats))
#####################################################################
Linear mixed-effects model fit by REML
 Data: Oats
  AIC BIC logLik
  603 614   -297

Random effects:
 Formula: ~Variety - 1 | Block
 Structure: Compound Symmetry
                  StdDev Corr
VarietyGolden Rain 18.2
VarietyMarvellous  18.2   0.635
VarietyVictory     18.2   0.635 0.635
Residual           12.9

Fixed effects: yield ~ nitro
            Value Std.Error DF t-value p-value
(Intercept)  81.9      6.95 65    11.8  <.0001
nitro        73.7      6.78 65    10.9  <.0001
 Correlation:
      (Intr)
nitro -0.293
```

5.4.2 Split-plot experiment on Oats - alternative way 2

Define the design matrix and random effects as

$$\mathbf{Z_i} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \mathbf{b_i}^\star = \begin{bmatrix} b_i \\ b_{i,1} \\ b_{i,2} \\ b_{i,3} \end{bmatrix}$$

and $\mathbf{b_i}^\star \sim N(0, \psi^\star)$. The covariance matrix for random effects is

$$\psi_{\mathbf{i}}{}^\star = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_2^2 & \\ & & & \sigma_2^2 \end{bmatrix}$$

the R code and output is

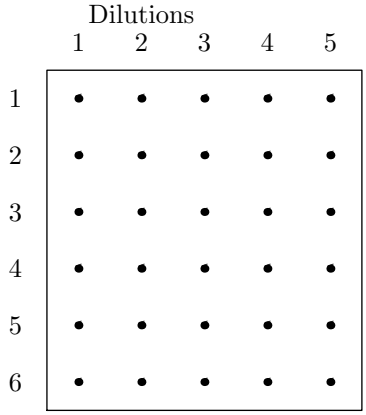
```
 mm2.Oats <- lme(yield ~ nitro,data=Oats,
  random=list(Block=pdBlocked(list(pdIdent( ~1),pdIdent(~Variety-1) ) )))
print(summary(mm2.Oats))
############################################################
Linear mixed-effects model fit by REML
 Data: Oats
  AIC BIC logLik
  603 614   -297

Random effects:    Composite Structure: Blocked
 Block 1: (Intercept)
 Formula: ~1 | Block
        (Intercept)
StdDev:        14.5
 Block 2: VarietyGolden Rain, VarietyMarvellous, VarietyVictory
 Formula: ~Variety - 1 | Block
 Structure: Multiple of an Identity
        VarietyGolden Rain VarietyMarvellous VarietyVictory Residual
StdDev:                 11                11             11     12.9

Fixed effects: yield ~ nitro
            Value Std.Error DF t-value p-value
(Intercept)  81.9      6.95 65    11.8  <.0001
nitro        73.7      6.78 65    10.9  <.0001
 Correlation:
      (Intr)
nitro -0.293
```

5.5 Crossed Random Effects

Crossed random effects are modelled by a combination of `pdBlocked` and `pdIdent` objects.

Example

The data are log(optical density) measures from cell cultures in 2 blocks of 30 wells,



There are 6 samples (treatments) randomly assigned to rows and 5 serial dilutions randomly assigned to columns. the data are `Assay` in the `nlme` library.

The systematic effects are `sample*dilut` and the random effects are `block/(row + column)`.

Index the blocks by i, the rows by j, columns by k. Then

$$y_{ijk} = \mu + s_j + d_k + (s:d)_{jk} + b_i + r_{ij} + c_{ik} + \epsilon_{ijk} \quad ,$$

$$\begin{array}{llll} b_i \sim N(0,\sigma_b^2) & r_{ij} \sim N(0,\sigma_r^2) & c_{ij} \sim N(0,\sigma_c^2) & \epsilon_{ijk} \sim N(0,\sigma^2) \\ & \mathrm{cov}(b,r) = 0 & \mathrm{cov}(b,c) = 0 & \mathrm{cov}(r,c) = 0 \end{array}$$

Write the vector of random effects as

$$\mathbf{U_i} = \begin{bmatrix} b_i \\ r_{i,1} \\ \vdots \\ r_{i,6} \\ c_{i,1} \\ \vdots \\ c_{i,5} \end{bmatrix} \quad .$$

Then

$$\mathrm{var}(\mathbf{U_i}) = \left[\begin{array}{c|c|c} \sigma_b^2 & & \\ \hline & \sigma_r^2\mathbf{I}_6 & \\ \hline & & \sigma_c^2\mathbf{I}_5 \end{array}\right] \quad .$$

That is $\mathbf{U_i}$ has a block-diagonal structure with each block being a multiple of the identity. The R code to construct this is

```
data(Assay)
mm.assay <- lme(logDens ~ sample*dilut, Assay,
   random=pdBlocked(list( pdIdent(~1),pdIdent(~ sample-1),pdIdent(~dilut-1))))
print(summary(mm.assay))
################################################
Linear mixed-effects model fit by REML
 Data: Assay
    AIC  BIC logLik
  -9.07 38.6   38.5

Random effects:
 Composite Structure: Blocked

 Block 1: (Intercept)
 Formula: ~1 | Block
        (Intercept)
StdDev:     0.00981

 Block 2: samplea, sampleb, samplec, sampled, samplee, samplef
 Formula: ~sample - 1 | Block
 Structure: Multiple of an Identity
        samplea sampleb samplec sampled samplee samplef
StdDev:  0.0253  0.0253  0.0253  0.0253  0.0253  0.0253

 Block 3: dilut1, dilut2, dilut3, dilut4, dilut5
 Formula: ~dilut - 1 | Block
 Structure: Multiple of an Identity
         dilut1  dilut2  dilut3  dilut4  dilut5 Residual
StdDev: 0.00914 0.00914 0.00914 0.00914 0.00914   0.0416

             numDF denDF F-value p-value
(Intercept)      1    29     538  <.0001
sample           5    29      11  <.0001
dilut            4    29     421  <.0001
sample:dilut    20    29       2   0.119
```

Chapter 6

Change Over designs

Designs in which each experimental unit receives a cyclical sequence of several treatments in successive periods are known as *change-over* designs. Historically they found favour because subject effects could be eliminated from experiment error but with the penalty that performance in a given period might reflect not only the direct effect of the treatment but also the residual effects of preceding treatments. Another reason for using change over designs is to get replication over periods when there is a shortage of experiment material and the researcher feels it is safe to get extra replication by giving more than 1 treatment to each experiment unit. Animal pen studies are sometimes done this way.

Direct and residual effects can be separated by appropriate choice of treatment sequences when it can be assumed that the residual effects persist only one period. We consider (i) balanced and (ii) partially balanced change over designs which may have an *extra period* which are formed by simply repeating the final period. The extra period provides the property that each treatment is preceded by the others equally often - ie a treatment is also preceded by itself. If first residuals are important, the extra period design is better but the extra period is unnecessary if residual effects are negligible.

In balanced designs, have all treatment contrasts of equal precision and in partially balanced designs, some contrasts are estimated with greater precision than others.

6.1 Latin squares and incomplete blocks

The notation is

- t - treatments
- p - periods
- b - blocks
- k - number of units per block

Table 6.1: Change-over design for 4 treatments using orthogonal Latin squares

	Square 1				Square 2				Square 3				Square 4			
	UNITS															
Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
2	2	1	4	3	3	4	1	2	4	3	2	1	2	4	1	3
3	3	4	1	2	4	3	2	1	2	1	4	3	3	1	4	2
4	4	3	2	1	2	1	4	3	3	4	1	2	4	3	2	1

Table 6.2: Change-over design for 7 treatments and 4 periods in blocks of 7

period	Incomplete Square 1							Incomplete Square 2						
1	1	2	3	4	5	6	7	1	2	3	4	5	6	7
2	2	3	4	5	6	7	1	7	1	2	3	4	5	6
3	4	5	6	7	1	2	3	5	6	7	1	2	3	4
4	7	1	2	3	4	5	6	2	3	4	5	6	7	1

The simplest type of change over design is a latin square with rows representing periods of time and columns representing experimental units. If there were no carry over effects, the data would simply be analysed as if it arose from an ordinary latin square. Residual effects are allowed for by including terms for them in the statistical model.

Complete sets of orthogonal Latin squares, eg Table 6.1 will ensure that each treatment is preceded by each other equally often but limits on resources will usually not allow this.

Many of the design properties are retained when rows are dropped from the latin square ($p < t$) or for an incomplete block where $k < t$. In an incomplete latin square, each treatment receives an incomplete set of treatments.

In order to estimate direct and residual effects, block size must be at east 3.

6.2 Analysis

If t_1, t_2, t_3 represent the direct effects of 3 treatments and r_1, r_2, r_3 the residual effects, the total effects of a sequence of treatments 1,2,3 are represented by

1. t_1
2. $t_2 + r_1$
3. $t_3 + r_2$

Thus the statistical model has components due to

1. Blocks (random effects)
2. Periods within Blocks (random effects)
3. Units within Blocks (random effects)
4. Direct effects (fixed effects)
5. Residual effects (fixed effects)

and we would regard items 1,2,3 as nuisance and items 4 and 5 as interest.

Example 6.1

The design is 4 blocks of orthogonal 3×3 Latin squares with an extra period where the rows are Periods and the columns are units.

	Block 1 Unit			Block 2 Unit			Block 3 Unit			Block 4 Unit		
Period	1	2	3	1	2	3	1	2	3	1	2	3
1	1	4	3	1	3	2	3	4	2	1	2	4
2	4	3	1	3	2	1	4	2	3	2	4	1
3	3	1	4	2	1	3	2	3	4	4	1	2
4	3	1	4	2	1	3	2	3	4	4	1	2

The data shown in Table 6.3 are milk yields (fcm) and the extra periods, indicated by †, could be dropped to compare the basic design with the extra period design.

Within each Block we have the additional blocking factors of Periods (the rows) and Units (the columns), the random model is

$$\text{Block} + \text{Block.Period} + \text{Block.Unit}$$

where *Block.Period* would require 3 df for each block, ie 12 df, and there would be 8 df for *Block.Unit*.

The systematic effects are the direct and residual treatment effects.

The statistical model for these data is

$$\begin{aligned} y_{itsu} &= \tau_i + \mathrm{R}_{i',(t-1)} + \mathrm{B}_s + \mathrm{B}_s.\mathrm{P}_t + \mathrm{B}_s.\mathrm{U}_u + \epsilon_{itsu} \\ & \mathrm{B}_s \sim N(0, \sigma^2_B) \;\; \mathrm{B}_s.\mathrm{P}_t \sim N(0, \sigma^2_{BP}) \;\; \mathrm{B}_s.\mathrm{U}_u \sim N(0, \sigma^2_{BU}) \;\; \epsilon \sim N(0, \sigma^2) \end{aligned}$$

To fit this model, we need to generate indicator variables R_2, R_3,R_4 to denote the treatment which preceded each particular case so that the coefficients of these indicator variables estimate the residual effect. The indicators for residual effects were plugged in by hand but could be generated with smart computing, eg

```
d1 <- dim(milk.g)[1]
units <- 1:d1
#      Calculate indicators for residual effects
milk.g$R4 <- milk.g$R3 <- milk.g$R2 <- milk.g$R1 <- rep(0,d1)
Tmat <- model.matrix( ~ Treat - 1,data=milk.g)
Pmat <- model.matrix( ~ Period -1,data=milk.g)
notP4 <- apply(Pmat[,1:3],1,sum)

for (i in 1:4){
Tind <- Tmat[,i]*notP4    # indicator of Ti occurring in periods 1, 2 or 3
Rirows <- units[Tind==1] + 1  # move 1 down
switch(i,
       milk.g[Rirows,"R1"] <- 1,
       milk.g[Rirows,"R2"] <- 1,
       milk.g[Rirows,"R3"] <- 1,
       milk.g[Rirows,"R4"] <- 1,
  stop() )
           }


milk.g <- groupedData(fcm ~1  | Block,data=milk)

milk.lme1 <- lme(fcm ~ Treat  + R2 + R3 + R4 -1,data=milk.g,
   random=pdBlocked(list(pdIdent( ~1),pdIdent(~Unit -1),pdIdent(~ Period -1))))
print(summary(milk.lme1)$tTable)     #$
```

The function pdIdent() produces a patterned matrix, Identity in this case.

The estimates of the direct and residual effects are:-

```
           Value  Std.Error
Treat1      27.4   3.5
Treat2      27.7   3.5
Treat3      27.9   3.5
Treat4      28.1   3.5
R2         -0.06   0.42
R3          0.25   0.42
R4         -0.06   0.42
```

Example 6.2

A Incomplete Block design for 6 treatments in 4 blocks, 4 periods and 4 units per block is shown below.

	Block 1				Block 2				Block 3			
	Unit				Unit				Unit			
Period	1	2	3	4	1	2	3	4	1	2	3	4
1	1	2	5	4	4	6	3	1	5	2	6	3
2	4	1	2	5	1	4	6	3	6	5	3	2
3	2	5	4	1	6	3	1	4	2	3	5	6
4	5	4	1	2	3	1	4	6	3	6	2	5

The model is the same as before and the essential R code is

```
XOIB.g <- groupedData(y ~1  | Block,data=XOIB)

XOIB.lme1 <- lme(y ~ Treat  + R2 + R3 + R4 + R5 + R6 -1,data=XOIB.g,
   random=pdBlocked(list(pdIdent( ~1),pdIdent(~Unit -1),pdIdent(~ Period -1))))
print(summary(XOIB.lme1)$tTable)                          #$
```

6.3 Change Over Designs - computing Guide

```
milk <- expand.grid(Period=1:4,Unit=1:3,Block=1:4)
milk$Treat <- c(1,4,3,3,4,3,1,1,3,1,4,4, 1,3,2,2,3,2,1,1,2,1,3,3,
                3,4,2,2,4,2,3,3,2,3,4,4, 1,2,4,4,2,4,1,1,4,1,2,2)
milk$fcm <- c(38.7,37.4,34.3,31.3,48.9,46.9,42.0,39.6,35.2,33.5,28.4,25.1,
              34.6,32.3,28.5,27.1,32.9,33.1,27.5,25.1,30.4,29.5,26.7,23.1,
              25.7,26.1,23.4,18.7,30.8,29.3,26.4,23.2,25.4,26.0,23.9,19.9,
              21.8,23.9,21.7,17.6,21.4,22.0,19.4,16.6,22.8,21.0,18.6,16.1)

milk$Subject <- factor((milk$Block-1)*3 + milk$Unit)
milk$Block <- factor(milk$Block)
milk$Period <- factor(milk$Period)
milk$Unit <- factor(milk$Unit)
milk$Treat <- factor(milk$Treat)

milk$R4 <- milk$R3 <- milk$R2 <- milk$R1 <- rep(0,48)
milk$R1[c(2,8,11,14,20,23,38,44,47)] <- 1
milk$R2[c(16,19,22,28,31,34,39,42,48)] <- 1
milk$R3[c(4,7,10,15,18,24,26,32,35)] <- 1
milk$R4[c(3,6,12,27,30,36,40,43,46)] <- 1

library(nlme)
milk.g <- groupedData(fcm ~1  | Block,milk)

milk.lme1 <- lme(fcm ~ Treat  + R2 + R3 + R4 -1,data=milk.g,
   random=pdBlocked(list(pdIdent( ~1),pdIdent(~Unit -1),pdIdent(~ Period -1))))
print(summary(milk.lme1)$tTable)    #$
```

Table 6.3: 4 treatments in a 3 × 3 latin square + extra period

Block	Unit	Subject	Period	Treat	R_1	R_2	R_3	R_4	fcm
1	1	1	1	1	0	0	0	0	38.7
1	1	1	2	4	1	0	0	0	37.4
1	1	1	3	3	0	0	0	1	34.3
1	1	1	4	3	0	0	1	0	31.3 †
1	2	2	1	4	0	0	0	0	48.9
1	2	2	2	3	0	0	0	1	46.9
1	2	2	3	1	0	0	1	0	42.0
1	2	2	4	1	1	0	0	0	39.6 †
1	3	3	1	3	0	0	0	0	35.2
1	3	3	2	1	0	0	1	0	33.5
1	3	3	3	4	1	0	0	0	28.4
1	3	3	4	4	0	0	0	1	25.1†
2	1	4	1	1	0	0	0	0	34.6
2	1	4	2	3	1	0	0	0	32.3
2	1	4	3	2	0	0	1	0	28.5
2	1	4	4	2	0	1	0	0	27.1 †
2	2	5	1	3	0	0	0	0	32.9
2	2	5	2	2	0	0	1	0	33.1
2	2	5	3	1	0	1	0	0	27.5
2	2	5	4	1	1	0	0	0	25.1 †
2	3	6	1	2	0	0	0	0	30.4
2	3	6	2	1	0	1	0	0	29.5
2	3	6	3	3	1	0	0	0	26.7
2	3	6	4	3	0	0	1	0	23.1 †
3	1	7	1	3	0	0	0	0	25.7
3	1	7	2	4	0	0	1	0	26.1
3	1	7	3	2	0	0	0	1	23.4
3	1	7	4	2	0	1	0	0	18.7 †
3	2	8	1	4	0	0	0	0	30.8
3	2	8	2	2	0	0	0	1	29.3
3	2	8	3	3	0	1	0	0	26.4
3	2	8	4	3	0	0	1	0	23.2 †
3	3	9	1	2	0	0	0	0	25.4
3	3	9	2	3	0	1	0	0	26.0
3	3	9	3	4	0	0	1	0	23.9
3	3	9	4	4	0	0	0	1	19.9 †
4	1	10	1	1	0	0	0	0	21.8
4	1	10	2	2	1	0	0	0	23.9
4	1	10	3	4	0	1	0	0	21.7
4	1	10	4	4	0	0	0	1	17.6 †
4	2	11	1	2	0	0	0	0	21.4
4	2	11	2	4	0	1	0	0	22.0
4	2	11	3	1	0	0	0	1	19.4
4	2	11	4	1	1	0	0	0	16.6 †
4	3	12	1	4	0	0	0	0	22.8
4	3	12	2	1	0	0	0	1	21.0
4	3	12	3	2	1	0	0	0	18.6

Chapter 7

Semi-parametric regression

A modern statistical concept is to use the flexibility of a smoother to represent the systematic effect and parametric densities to model the random effects. If we categorize the model by parameters $\theta \in \Omega$, then for particular types of models,

parametric	Ω is finite dimensional Euclidean space,
semi-parametric	Ω is a combination of infinite dimensional space and finite dimensional Euclidean space,
non-parametric	Ω is infinite dimensional space

One must keep this in mind when interpreting the models.
Consider categories of smoothers,

B-splines	kernel	splines
Wavelets	loess	
	projection pursuit	

Splines are often the smoother of choice because their flexibility in modelling situations but with mathematics, it is often possible to transform one form of smoother into another. Wavelets are almost solely the preserve of high intensity data such as in signals and images.

7.1 Generalized Cross validation

Suppose

$$y = g(\eta) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \text{ where} \tag{7.1}$$
$$\eta = \beta_0 + s_1(x_1) + s_2(x_2) + \ldots \tag{7.2}$$

The smooth terms can be written as a suitable set of basis functions,

$$s(x) = \sum_{j=1}^{k} \beta_j b_j(x)$$

where the $b_j(x)$ are a set of basis functions such as

- Hermite polynomials,

$$b_0(x) = 1, \quad b_1(x) = 2x, \quad b_2(x) = 4x^2 - 2, \; b_3(x) = 8x^3 - 12x,$$
$$b_n(x) = e^{2nx - n^2}, \quad \int_{-1}^{1} b_m(x) b_n(x) = 0$$

- cubic splines

$$b_1 = x \quad , b_j = |x - x_j^*|^3 \quad , j > 1$$

where x_j^* are the knot points.

Hence (7.2) can be written as a linear model and in the case of cubic splines,

$$\begin{aligned} \eta &= \beta_0 + \beta_1 x - 1 + \sum_{j=1}^{k_1} \beta_{j+1} |x_1 - x_{1,j}^*| + \ldots \qquad \text{or} \\ \eta &= X\beta \end{aligned}$$

At this point the degrees of freedom depend on the number of knot points . To avoid influence of the knot points, the model overfits and controls smoothness by a penalty function. So it may be considered as putting in a lot of df and then taking out the redundant ones *where the data indicate they are not needed*. Consequently we may end up with fractional df.

It is timely to note the Littlewood-Paley-Stein theorem,

Theorem 1 *If $1 < p < \infty$, there exist 2 constants, $C_p \geq c_p > 0$ such that for all functions belonging to $L^p(\mathcal{R}^n)$,*

$$c_p ||g||_p \leq ||f_p|| \leq C_p ||g||_p \quad ,$$

where

$$||f||_p = \left(\int_{\mathcal{R}^n} |f(x)|^p \right)^{\frac{1}{p}}$$

and

$$g(x) = \left(\sum_{x=-\infty}^{\infty} | \sum_{2^j \leq k \leq 2^{j+1}} (a_k \cos kx + b_k \sin kx)|^2 \right)^{\frac{1}{2}}$$

The term $g(x)$ is the Fourier transform of the data y and this theorem is stating properties of the basis functions used to represent y. With well chosen basis functions, c_p and C_p will be close $\forall p$. With poorly chosen basis functions, the addition of extra terms is not so much fine tuning as fixing the discrepancies of the low order terms and so the representation of y by $\sum b_i(x)$ has localized hot spots of spectral energies. If the spectral energy is distributed evenly across x, $||f||_p$ will be approximately the same for $p = 2, 4, 6, \ldots$.

The penalty function for splines is the roughness defined by

$$\int [s''(x)]^2 \, dx = \beta^T S \beta$$

and the β's are estimated by minimizing

$$-\ell(\beta) + \sum_i \lambda_i \beta^T S \beta \quad \text{where } \lambda_i \text{ are weights.}$$

In the GLM setting we require $\Gamma_{i,i} = \frac{\partial y_i}{\partial \mu_i}$ and $W_{i,i} = \left(\frac{\partial \mu_i}{\partial y_i}\right) / \frac{\partial \mu_i}{\partial \theta_i}$, θ being the canonical parameter. Then the steps in fitting a semiparametric model are:-

1.
$$z = \eta + \Gamma(y - \mu)$$

2. find λ_i that minimizes
$$\frac{||W^{\frac{1}{2}}(z - X\beta||^2}{\text{tr}(I - A)^2}$$
where $A = X(X^T W X + \sum_i \lambda_i \beta^T S \beta)^{-1} W$ is the hat matrix and $\text{tr}(A) = \text{df}$.

This algorithm is known as generalized cross validation.

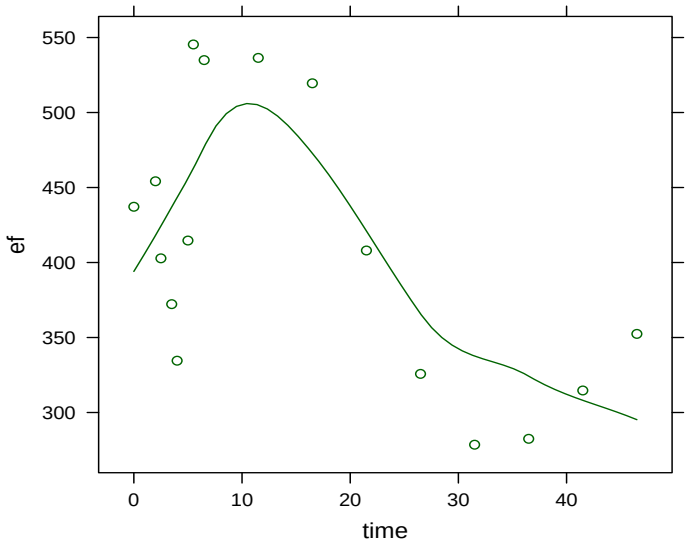
7.2 GAM's in R

The mgcv library in R[1] fits Generalized Additive Models using generalized cross validation. The example is for modelling elbow flux over time. A subject receives some physiotherapy treatment and the blood flow through the elbow is measured (as elbow flux) for the next hour. In the study, there were many subjects but only the data from one is used here to demonstrate GAMS. The data are saved in `elbowdat.csv` and an exploratory data analysis shows a highly nonlinear trend of `ef` over `time`.

```
library(mgcv);library(lattice)
elbow.df <- read.csv("elbowdat.csv",header=T)
ef.eda <- xyplot(ef ~ time,data=elbow.df,type=c('p','smooth'))
print(ef.eda)
```

[1] `fttp://mirror.aarnet.edu.au/pub/CRAN`

```
time,   ef
   0, 437.18
   2, 454.17
 2.5, 402.79
 3.5, 372.25
   4, 334.55
   5, 414.66
 5.5, 545.38
 6.5, 534.95
11.5, 536.43
16.5, 519.49
21.5, 408.01
26.5, 325.76
31.5, 278.54
36.5, 282.47
41.5, 314.69
46.5, 352.42
```



There is slight evidence that the variance is proportional to the mean although the cluster of points between 10 and 20 minutes which are approximately ef= 530, would not support that assertion. Nevertheless, over the whole data set (trust me), the gamma distribution was deemed to adequately model the random component.

```
model1 <- gam(ef ~ s(time),data=elbow1,family=Gamma(link=log))
```

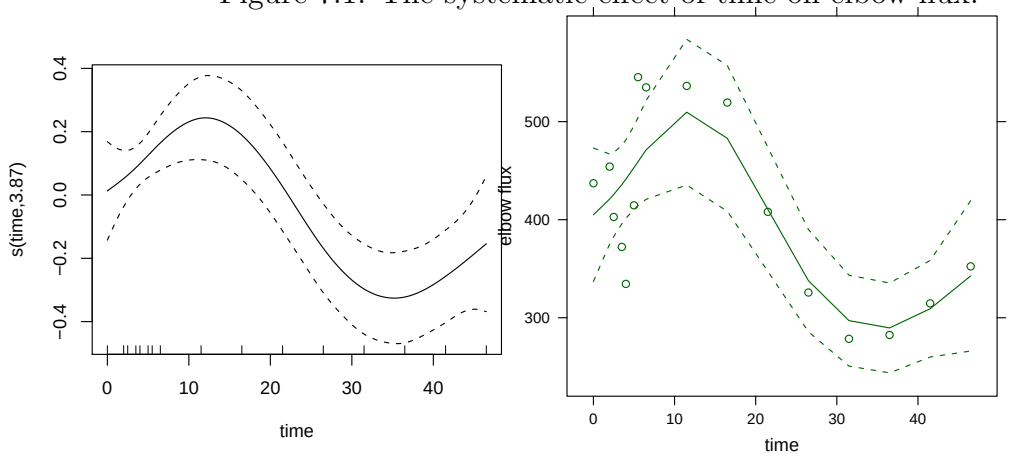
```
> anova(model1)
Family: Gamma
Link function: log

Formula:  ef ~ s(time)
Approximate significance of smooth terms:
          edf Est.rank     F p-value
s(time) 3.868    8.000 5.104 0.00736
```

The AOV indicates that the spline satisfactorily represents the systematic effect.

The fit of the spline component is shown in Figure 7.1. In the left frame, the plot is of the fitted trend on the log scale of the link function. The right frame shows the fitted response with 95% CI's..

```
plot(model1)

preds <- predict(model1,se=T,type="response")
elbow.df$fitted <- preds$fit
elbow.df$lwr <- preds$fit - 2* preds$se.fit
elbow.df$upr <- preds$fit + 2* preds$se.fit

trellis.device(theme="col.whitebg",width=5,height=4.5)
```

Figure 7.1: The systematic effect of time on elbow flux.




```
fit.plot <- xyplot(ef+fitted+lwr+upr ~ time,data=elbow.df,type=c('p','l','l','l'),
ylab="elbow flux",
  panel=function(x,y){
  ntimes <- length(x)/4
l1 <- 1:ntimes
l2 <- (ntimes+1):(2*ntimes)
l3 <- (2*ntimes+1):(3*ntimes)
l4 <- (3*ntimes+1):(4*ntimes)
panel.xyplot(x[l1],y[l1])
panel.xyplot(x[l2],y[l2],type='l')
panel.xyplot(x[l3],y[l3],type='l',lty=2)
panel.xyplot(x[l4],y[l4],type='l',lty=2)
                        }
)
print(fit.plot)
dev.off()
```

The R newsletter June 2001 at the CRAN web site has a paper by Simon Wood on GAM's.

Example

The next example shows how to fit GAMs within a group such as the levels of a treatment.

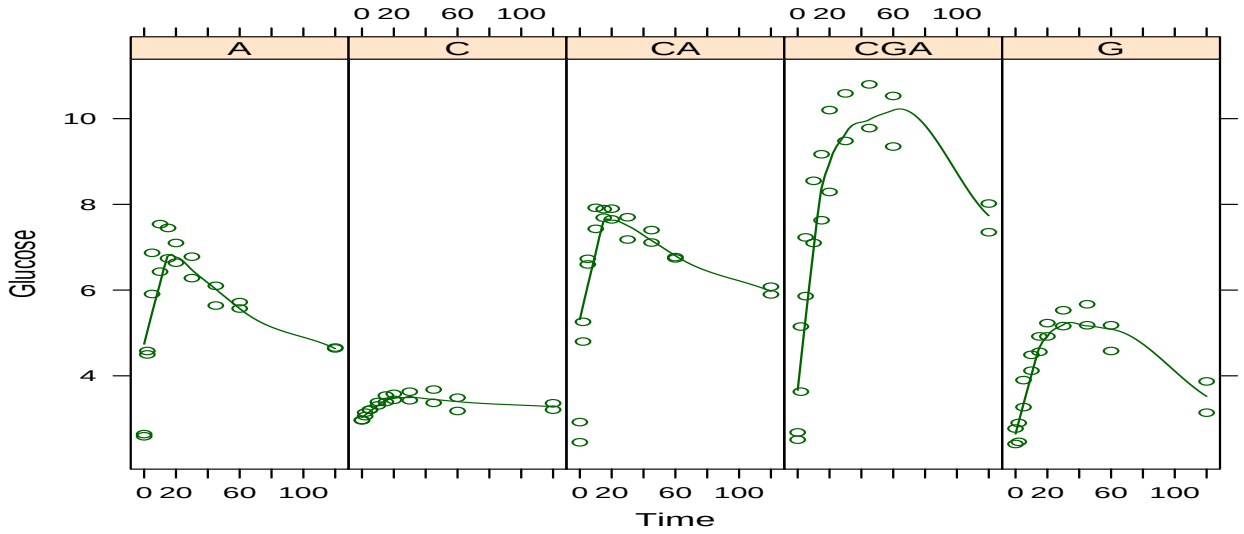
The data are `Glucose` concentrations in sheep taken over about 2 hours after the sheep had been injected with hormones,
(a) Cortisol (C), (b) Glucagon (G), (c) Adrenalin (A), (d) C+A, (e) C+G+A.
The full data are on the web site and only the first few lines are shown here.

```
Sheep Weight Treatment Time Glucose
148 32 C 0 2.97
148 32 C 2 3.06
148 32 C 5 3.21
148 32 C 10 3.31
148 32 C 15 3.39
148 32 C 20 3.44
```

```
library(mgcv)
Glucose.df <- read.table("Glucose.txt",header=T)

trellis.device(theme="col.whitebg")
GlucoseEDA.plot <- xyplot(Glucose ~ Time|Treatment,data=Glucose.df,
                    type=c('p','smooth'),layout=c(5,1)        )
print(GlucoseEDA.plot)
```

Exploratory data plots show that the responses of `Glucose` over time are non-linear and that the shape of response differs amongst treatments.

Figure 7.2: Exploratory data plots for Glucose response to hormones



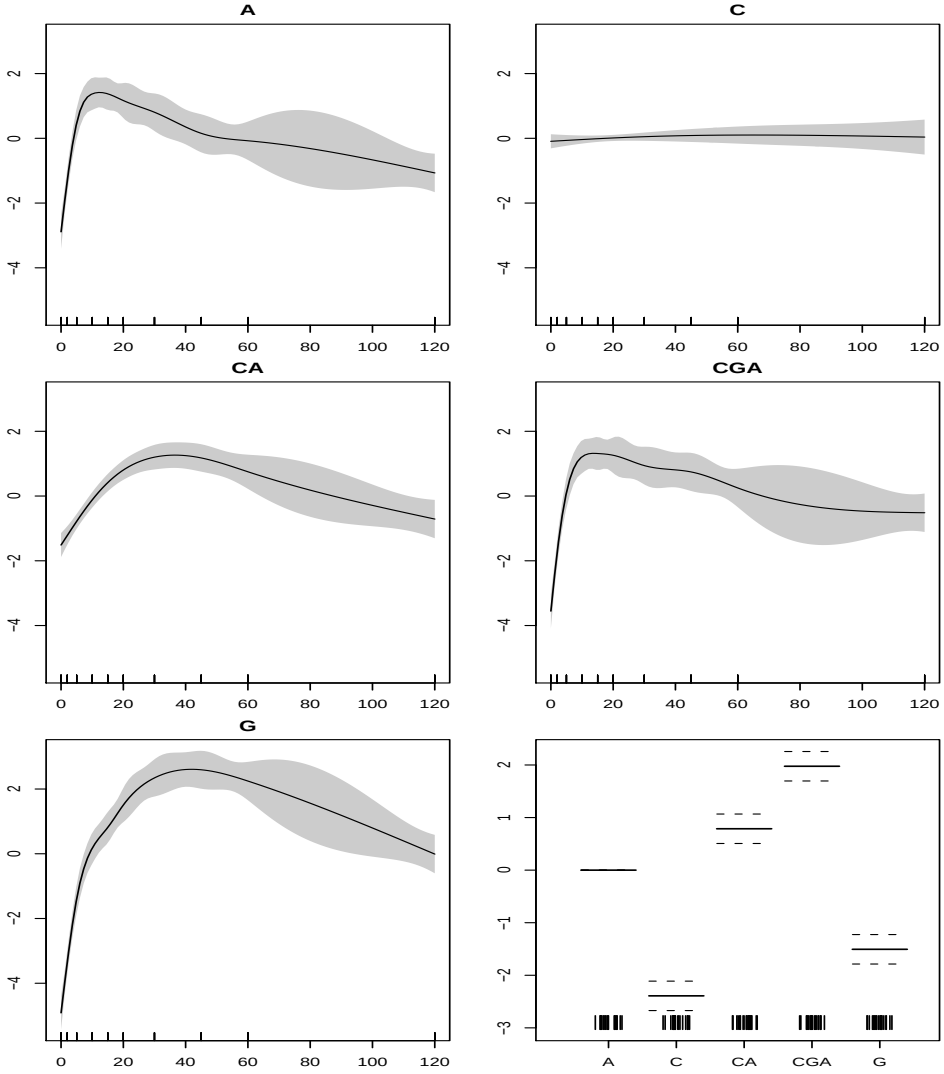
The concept of an interaction does not apply in semi-parametric regression and the effect of time is estimated within each treatment. This requires that we set up an indicator variable for each treatment and use this to specify that the spline fit is within that group.

```
trt.levels <- levels(Glucose.df$Treatment)
#  a loop to get dummy variables that indicate which treatment
     for (i in seq(along=trt.levels)){
          Glucose.df[,trt.levels[i]] <- as.numeric(Glucose.df$Treatment==trt.levels[i])
                                       }   #  end of the i loop

mod1 <- gam(Glucose ~ Treatment +
      s(Time,by=A)+s(Time,by=C)+s(Time,by=G)+s(Time,by=CA)+s(Time,by=CGA),
        data=Glucose.df )
print(anova(mod1))

par(mar=rep(2,4),mfrow=c(3,2) )
for(j in 1:5){
plot(mod1,shade=T,select=j,xlab="",ylab="",main=trt.levels[j])
              }
plot(mod1,all.terms=T,select=6)
```

The following code plots the fitted responses.

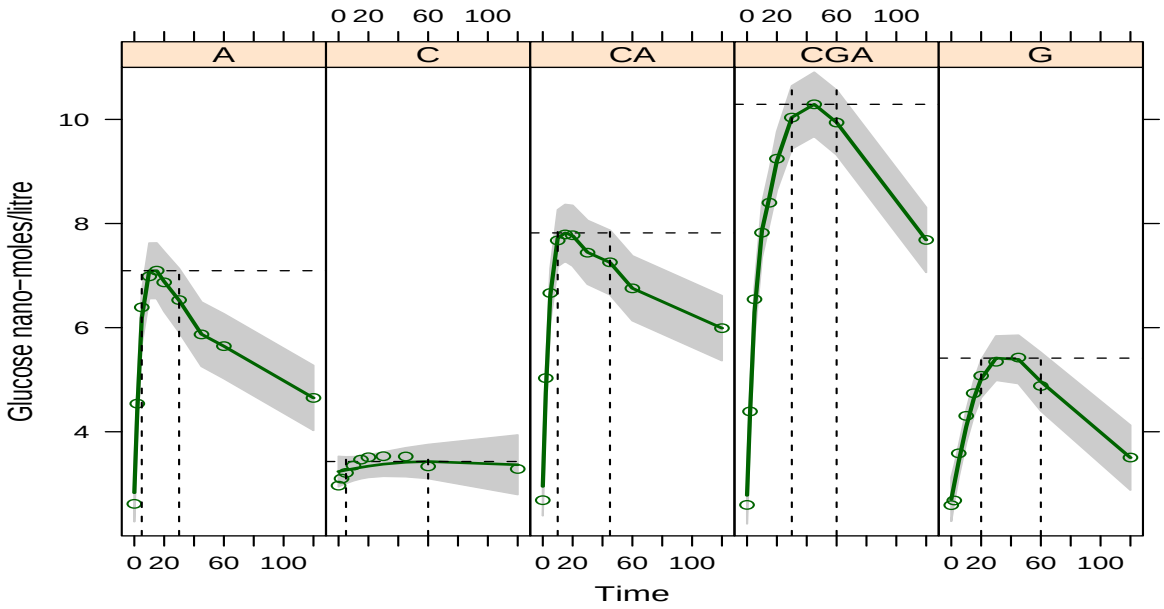
The plot of the model in Figure 7.3 shows the spline *effects* for each level and the `all.terms` option gives the comparison of mean values.

Figure 7.3: Fitted responses from GAM models within treatment levels



The profile of fitted values is used to estimate, with 95% CI's, the time to max concentration.

```
#_____________________________________________
closest <- function(s,v){   # a function to be used in the panel
delta <- abs(s-v)
pp <- order(delta)[1]
return(pp)}
#_____________________________________________
uTimes <- sort(unique(Glucose.df$Time))
ntimes <- length(uTimes)
pred.df <- expand.grid(Time=uTimes,Treatment=trt.levels)
for (i in seq(along=trt.levels)){
pred.df[,trt.levels[i]] <- as.numeric(pred.df$Treatment==trt.levels[i])
                                }   #  end of the i loop
preds <- predict(mod1,newdata=pred.df,se=T)
pred.df$fit <- preds$fit
pred.df$lwr <- preds$fit - 2*preds$se.fit
pred.df$upr <- preds$fit + 2*preds$se.fit

points.df <- aggregate(Glucose.df[,"Glucose"],
  by=list(Glucose.df$Time,Glucose.df$Treatment),FUN=mean)
pred.df$means <- points.df$x

ylim=c(2,11)
trellis.device(theme="col.whitebg",device="pdf",file="GlucoseFit.pdf",width=6,height=5)
fit.plot <- xyplot(fit+lwr+upr+means ~ Time|Treatment,data=pred.df,ylim=ylim,
   layout=c(5,1),ylab="Glucose nano-moles/litre",
  panel=function(x,y){
  ntimes <- length(x)/4
l1 <- 1:ntimes
l2 <- (ntimes+1):(2*ntimes)
l3 <- (2*ntimes+1):(3*ntimes)
l4 <- (3*ntimes+1):(4*ntimes)

  v <- c(x[l2],rev(x[l3]),x[l2[1]])
  w <- c(y[l2],rev(y[l3]),y[l2[1]])
  panel.polygon(v,w,col=gray(0.8),border=gray(0.8))
   panel.xyplot(x[l1],y[l1],type='l',lwd=2)
 panel.abline(h=max(y[l1]),lty=2)
 panel.points(x[l4],y[l4])

 xmax <- (l1)[y[l1]==max(y[l1])]
 x1 <- x[(2*ntimes+1):(2*ntimes+xmax)][closest(max(y[l1]),y[(2*ntimes+1):(2*ntimes+xmax)])]
 x2 <- x[(2*ntimes+xmax)  :(3*ntimes)][closest(max(y[l1]),y[(2*ntimes+xmax):(3*ntimes)])]
 y1 <- ylim[1]
 y2 <- y[l3][closest(max(y[l1]),y[l3]) ]
 panel.segments(x1,y1,x1,y2,lty=2)
 panel.segments(x2,y1,x2,y2,lty=2)
                      }
)
print(fit.plot)
dev.off()
```



7.3 Smoothing splines are BLUPS

In his comment upon Geoff Robinson's 1991 paper, *That BLUP is a Good Thing: The estimation of Random Effects* [8], Terry Speed [11] showed that splines are BLUPS. This permits the non-linear smooth splines to be implemented in a linear mixed model which substantial power for modelling data. This is expertly developed in Verbyla *et al.* (1999) and the models can be fitted using the specialized package *ASREML* written by Arthur Gilmour.

Theory and implementation are also given in the book "Semi-parametric Regression" by David Ruppert, Matt Wand and Ray Carroll [9].

Consider a linear model

$$y = g(x) + \epsilon \ , \quad \epsilon \sim N(0, \sigma^2 \mathbf{R})$$

and write -2 log-likelihood as

$$\ell = -2 \log |\sigma^2 \mathbf{R}| - \frac{1}{2\sigma^2} \left[(\mathbf{y} - \mathbf{g})^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{g}) + \lambda_s \int \left\{ g''(x) \right\}^2 dx \right]$$

The solution through REML estimation of variance components and mixed model estimating equations utilizes matrices which are functions of difference between samplings of the variable in the spline function. Define these differences as $h_j = x_{j+1} - x_j$. Then

$$\Delta = \begin{bmatrix} \ddots & & & & \\ \ddots & \ddots & & & \\ \ddots & \ddots & & \ddots & \\ & \frac{1}{h_{i+1}} & -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right) & \frac{1}{h_i} & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}_{n,n-2} ,$$

and

$$G = \begin{bmatrix} \ddots & \ddots & & & \\ \ddots & \ddots & \ddots & & \\ & \frac{h_{i+1}}{6} & \frac{h_i+h_{i+1}}{3} & \frac{h_{i+1}}{6} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}_{n-2,n-2}$$

At the design points,

$$\hat{\mathbf{g}} = \left(\mathbf{R}^{-1} + \lambda_s \Delta G^{-1} \Delta\right) \mathbf{R}^{-1} y \tag{7.3}$$

Equation (7.3) can be reparameterized into the standard form for a mixed model by utilizing,

$$\begin{aligned} X_s &= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\ Z_s &= \Delta(\Delta^T\Delta)^{-1} \\ H &= \sigma^2 \left(R + \lambda_s^{-1} Z_s G Z_s^T\right) \end{aligned}$$

so that

$$\hat{g} = X_s\hat{\beta}_s + Z_s\tilde{u}_s$$

$$\begin{aligned} \text{where}\hat{\beta}_s &= (X_s^T H^{-1} X_s)^{-1} X_s^T H^{-1} y \\ \tilde{u}_s &= (Z_s^T R^{-1} Z_s + \lambda_s G^{-1}) Z_s^T R^{-1} (y - X\hat{\beta}_s) \\ \lambda_s &= \frac{\sigma^2}{\sigma_s^2} = \gamma_s^{-1} \end{aligned}$$

The log-likelihood is a conditional likelihood of $y|u_s$ and the penalty is the log-density function of u_s. Splines in GLM's are fitted using the method of Schall (1991) [10].

7.4 Splines in the GLM using ASREML

The ASREML program [4] is not simple but its power makes it worth having and learning. The routines are in GENSTAT and a S-PLUS module is being fine tuned (I think). I use a primitive R function called asr() which sets the job up and after some translation, passes it to the compiled ASREML program. My R code is

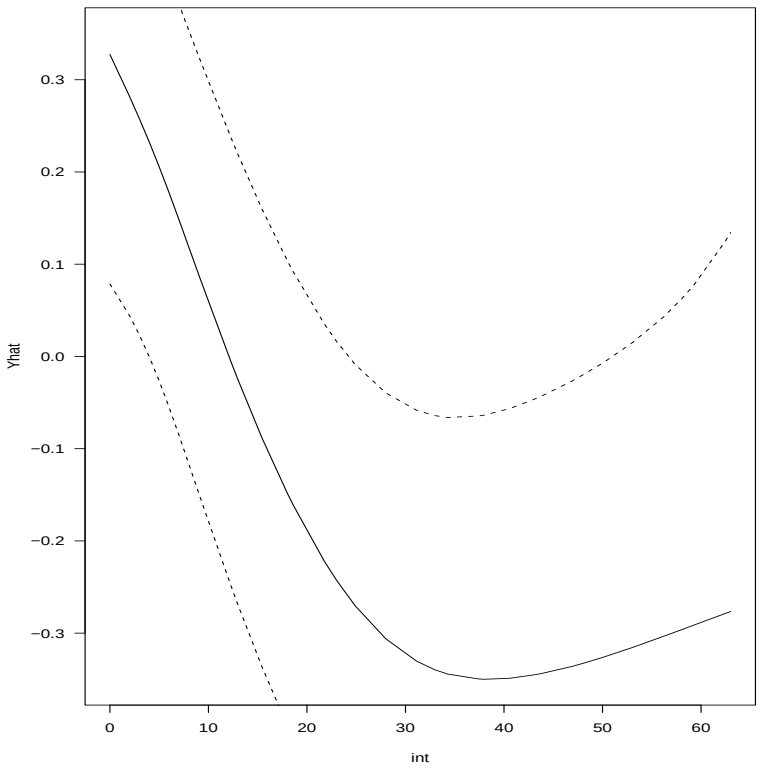
```
attach("/ASREML/.RData",pos=2)
model2 <- asr("ef !GAMMA !LOG ~ mu int !r spl(int) SUBNO SUBNO.int",data=elbow,title="elbow flu
detach(2)
```

which writes out the code for ASREML in a file called **elbow.as** and also rewrites the data into **elbow.asd**,

```
elbow flux
  SUBNO 25  !A
  TRT 19  !A
  int
  treatment 1  !A
  aet
  uet
  aht
  uht
  asc
  usc
  ef
  hf
elbow.asd    !skip 1
ef !GAMMA !LOG ~ mu int !r spl(int) SUBNO SUBNO.int
```

The fit is similar to that with gam() in R

Figure 7.4: Smooth splines for a Gamma model in ASREML



Chapter 8

Longitudinal Data

Repeated Measures data are those which consitute a set of repeated measures, over time, from the same sampling unit. Because the sampling is from the same unit, the data contains information from the unit (ie random effect) and the systematic effect of time. In conjunction with this, the measurements are correlated because they all have something in common, the unit itself. Another potential source of correlation is that an observation is influenced by previous observations and this is termed autoregressive correlations.

Modelling of repeated measures data needs to account for the correlations so that the model can best allocate the information between systematic and random effects.

8.1 The data file

Early repeated measures models were cast as multivariate models in order to capture the correlations amongst the data. This idea was superseded and now we analyze repeated measures in a univariate fashion,

$$\begin{bmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ \vdots \\ \mathbf{Y}_\tau \end{bmatrix} = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_\tau \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \epsilon \ .$$

We have the time variable as a column of the X_i's,

$$X_i = \begin{bmatrix} 1 & 0 & \dots & 1 & \dots & t_{i,1} \\ 1 & 0 & \dots & 1 & \dots & t_{i,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 1 & \dots & t_{i,n_i} \end{bmatrix}$$

The important difference between this model and simple linear models is that elements of ϵ are not assumed to be independent or to have constant variance,

An example of data file on carbon form Tillage and Rotation treatments, measured at 0,3,4,6,12 months is:-

Plot	Sample	Rotation	Tillage	Block	Time	C
P1	S1	R1	T1	B1	0	1.53
P2	S2	R1	T1	B1	0	1.47
P3	S3	R1	T1	B1	0	1.29
.	.	.	.	.	.	.
P9	S3	R2	T2	B1	0	2.06
P10	S1	R1	T1	B1	0	1.4
P11	S2	R2	T1	B1	0	1.46
P12	S3	R2	T1	B1	0	1.53
P1	S1	R1	T1	B1	3	1.21
P2	S2	R1	T1	B1	3	1.17
P3	S3	R1	T1	B1	3	1.31
.	.	.	.	.	.	.
P22	S1	R1	T2	B2	3	1.39
P23	S2	R1	T2	B2	3	1.91
P24	S3	R1	T2	B2	3	1.42
P1	S1	R1	T1	B1	4	1.04
P2	S2	R1	T1	B1	4	1.12
P3	S3	R1	T1	B1	4	1.07
.	.	.	.	.	.	.
P22	S1	R1	T2	B2	4	1.25
P23	S2	R1	T2	B2	4	1.21
P24	S3	R1	T2	B2	4	1.23
P1	S1	R1	T1	B1	6	1.44
P2	S2	R1	T1	B1	6	1.39
P3	S3	R1	T1	B1	6	1.15
.	.	.	.	.	.	.
P10	S1	R2	T1	B1	6	1.23
P11	S2	R2	T1	B1	6	1.17
P13	S1	R2	T1	B2	6	1.29
P14	S2	R2	T1	B2	6	1.14
.	.	.	.	.	.	.
P22	S1	R1	T2	B2	12	1.83
P23	S2	R1	T2	B2	12	1.70
P24	S3	R1	T2	B2	12	1.40

In this case all treatments are sampled at the same times 0,3,4,6 and 12 months and the data have been blocked for each sample time. Note that Plot=P12 at Time=6 is missing so that data is either deleted or denoted as a missing value NA.

The next example is measurements of Metabolizable Energy of deer which are born at different times and hence are sampled on different days. There are 25 animals and the data are blocked as all the repeated measurements for the first animal, followed by the repeated measurements for the second and so on.

Tag	Strain	Sex	Age	Agewks	ME	Lwt
204	3	1	227	32.43	123.22	71.50
204	3	1	234	33.43	125.56	73.00
204	3	1	241	34.43	122.94	75.00
204	3	1	248	35.43	123.00	74.50
204	3	1	255	36.43	117.83	75.00
204	3	1	262	37.43	107.25	75.50
204	3	1	269	38.43	108.58	75.00
.	.	.	.	.	.	.
207	3	1	224	32.00	133.05	67.50
207	3	1	231	33.00	131.76	67.50
207	3	1	238	34.00	137.45	67.50
207	3	1	245	35.00	128.62	68.00
207	3	1	252	36.00	129.41	69.00
207	3	1	259	37.00	129.24	69.00
.	.	.	.	.	.	.
293	4	2	540	77.14	108.91	96.50
293	4	2	547	78.14	151.51	98.50
293	4	2	554	79.14	131.97	98.50
293	4	2	563	80.43	125.17	99.50
293	4	2	568	81.14	138.74	101.00
293	4	2	575	82.14	128.27	96.50
293	4	2	582	83.14	121.92	99.50

If there are repeated measurements in two directions, eg time and depth, the data file is set up similar to the above but with both classifying variables, eg

Hole	Treatment	Stratum	Depth	Calendar	WC
E2B	Control	Beneath	20	07/08/95	19.62
E2B	Control	Beneath	20	07/25/95	18.98
.	.	.	.	.	.
E2B	Control	Beneath	40	07/08/95	17.64
E2B	Control	Beneath	40	07/25/95	17.74
.	.	.	.	.	.
E2B	Control	Beneath	60	07/08/95	16.97
E2B	Control	Beneath	60	07/25/95	16.86
.	.	.	.	.	.
E2B	Control	Beneath	80	07/08/95	17.88
E2B	Control	Beneath	80	07/25/95	17.92
.	.	.	.	.	.
E2B	Control	Beneath	100	07/08/95	11.8
E2B	Control	Beneath	100	07/25/95	11.5
.	.	.	.	.	.
E2B	Control	Beneath	120	07/08/95	9.41
E2B	Control	Beneath	120	07/25/95	9.41
.	.	.	.	.	.
F1B	Deforest	Beneath	20	01/27/94	16.0
F1B	Deforest	Beneath	20	02/11/94	14.70
.	.	.	.	.	.
F1B	Deforest	Beneath	40	01/27/94	17.30
F1B	Deforest	Beneath	40	02/11/94	14.91
.	.	.	.	.	.
F1B	Deforest	Beneath	60	01/27/94	18.54
F1B	Deforest	Beneath	60	02/11/94	14.40
.	.	.	.	.	.
F1B	Deforest	Beneath	80	01/27/94	19.82
F1B	Deforest	Beneath	80	02/11/94	14.86
.	.	.	.	.	.
F1B	Deforest	Beneath	100	01/27/94	23.69
F1B	Deforest	Beneath	100	02/11/94	18.65
.	.	.	.	.	.
F1B	Deforest	Beneath	120	01/27/94	23.86
F1B	Deforest	Beneath	120	02/11/94	23.69
.	.	.	.	.	.
F1B	Deforest	Beneath	140	01/27/94	20.10
F1B	Deforest	Beneath	140	02/11/94	21.63

The data are unbalanced in that sampling times are different for sites and the numbers of depths differ between treatments. The imbalance does not complicate the analysis.

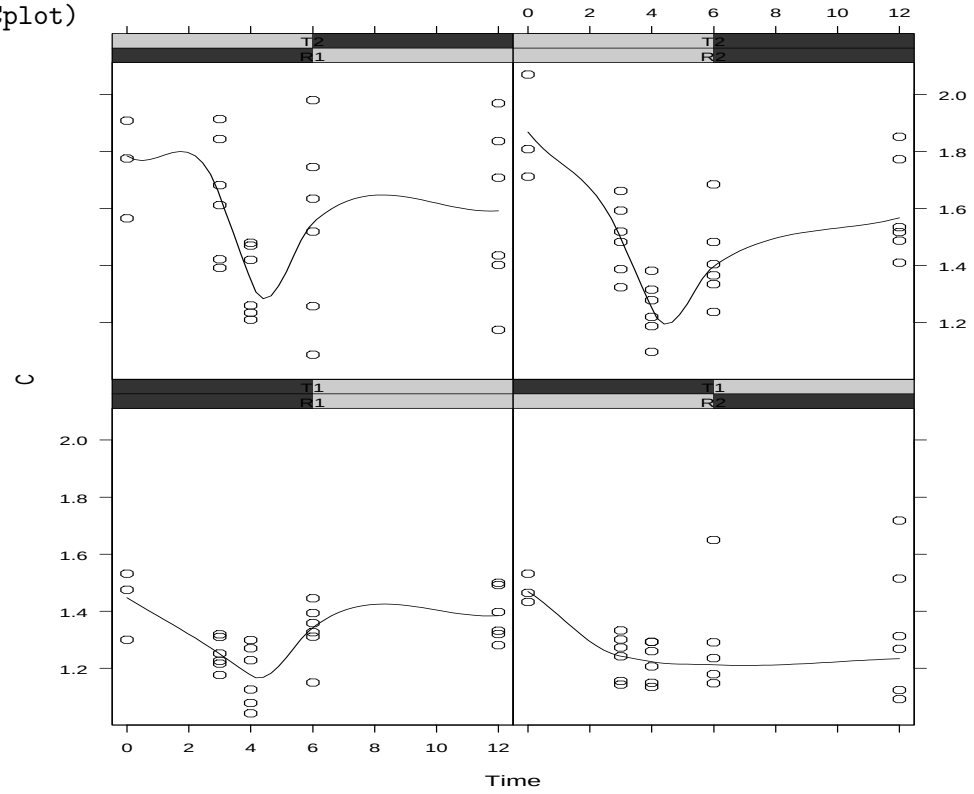
8.2 Exploratory Data Analysis

Analyses of longitudinal data compare profiles over time and indeed time might be viewed as the primary systematic effect to be investigated.

Lattice graphics in R can be tweaked to give multi-panel plots to reveal structure.

This example concerns repeated measures of carbon from 2 tillage × 2 rotation treatments.

```
Cplot_xyplot(C ~ Time| Rotation*Tillage,data=carbon,
      panel=function(x,y){panel.xyplot(x,y);panel.loess(x,y)})
print(Cplot)
```



This is not an analysis of the data but it does reveal the interaction of treatments and time and the changing variance. If we were to ignore the variance structure, it is as likely that the interaction would be missed because of the inappropriate allocation of the information to systematic and random effects.

8.3 Statistical model for repeated measures

The correlations and changing variances which are charactersitic of longitudinal data are often induced by the units such as animals or plots. This is because over time, observations within a unit may be more alike than those of the same treatment but sampled from another unit. Hence the random component is affected by sampling units.

The basic model is

$$y = \text{treatment} + \text{time} + \text{treatment} : \text{time} + \text{ERROR} \quad , \text{ERROR} \sim \text{A distribution}$$

Note how general this is and the error model is not restricted to a simplistic (and unrealistic) i.i.d. Momentarily we regard both treatment and time as factors for which we need to construct contrasts to get a more specific model. Also, we have recognised that the random structure may contain a number of model terms.

The model may then contain these components,

$$\begin{aligned} y \quad = \quad & \text{treatment} + \text{time} + \text{treatment} : \text{time} + \text{unit} + \text{unit} : \text{time} + \text{error} \\ & \\ & \text{unit} \sim N(0, \sigma_u^2 \times J_1) \\ & \text{unit} : \text{time} \sim N(0, \sigma_{ut}^2 \times J_2) \\ & \text{error} \sim N(0, \sigma^2 \times R) \end{aligned}$$

The symbols J_1, J_2, R denote matrices which give the weights to be applied to the variance components for each part of the model.

That is the model for the random components may include,

- Measurement error,
- Components due to random effects,
- Serial correlation

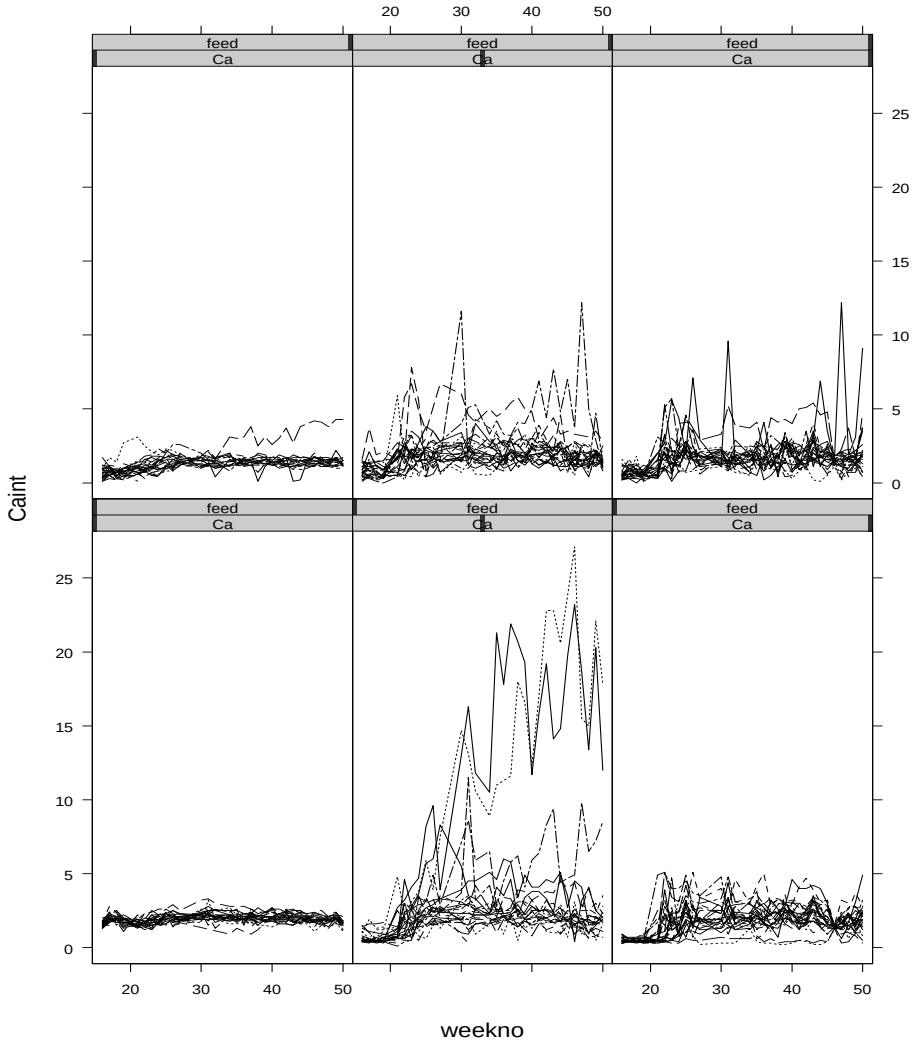
8.3.1 Random components

A line plot of the response over time for each unit can sometimes tell us about the random component. If the plot reveals basically parallell lines but with different intercepts, the appropriate model regards the intercept as a random effect. In this case, the correlations amongst the repeated measures are uniform.

The more usual case is that the lines fan out because the individuals change differently over time. In that case, the interaction of the individual and the systematic time effect would be fitted as a random effect and this representation leads to a covariance structure that is dependent upon the differences between sampling times.

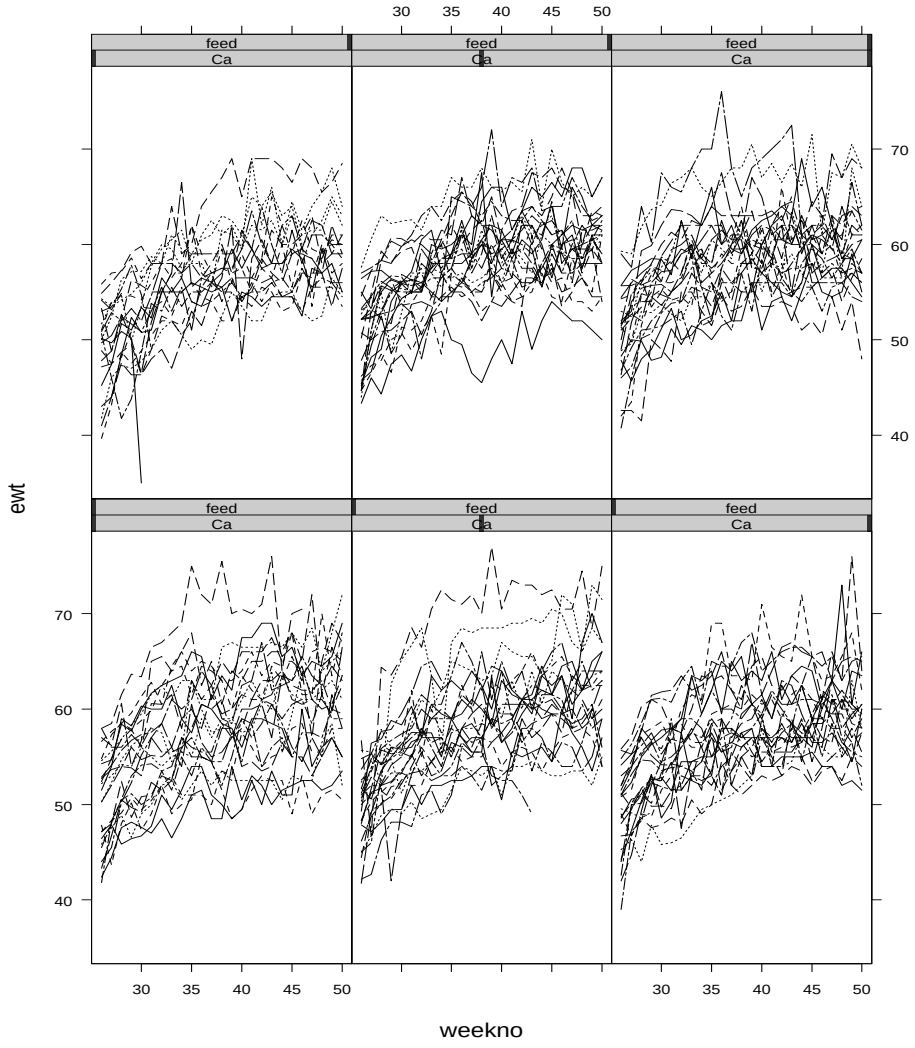
The following 3 examples are (i) Ca intake, (ii) egg weight measured on the same birds and the ME of deer. The Ca intake data show a fan out over time suggestion and animal $\times$ time interaction and the egg weight data suggest that differences are primarily due to animals only or a random intercept model. The deer data show a fanning out at the maximum.

```
xyplot(Caint ~ weekno|Ca*feed,type='l',data=Caint,
panel=panel.superpose,groups=an)
```



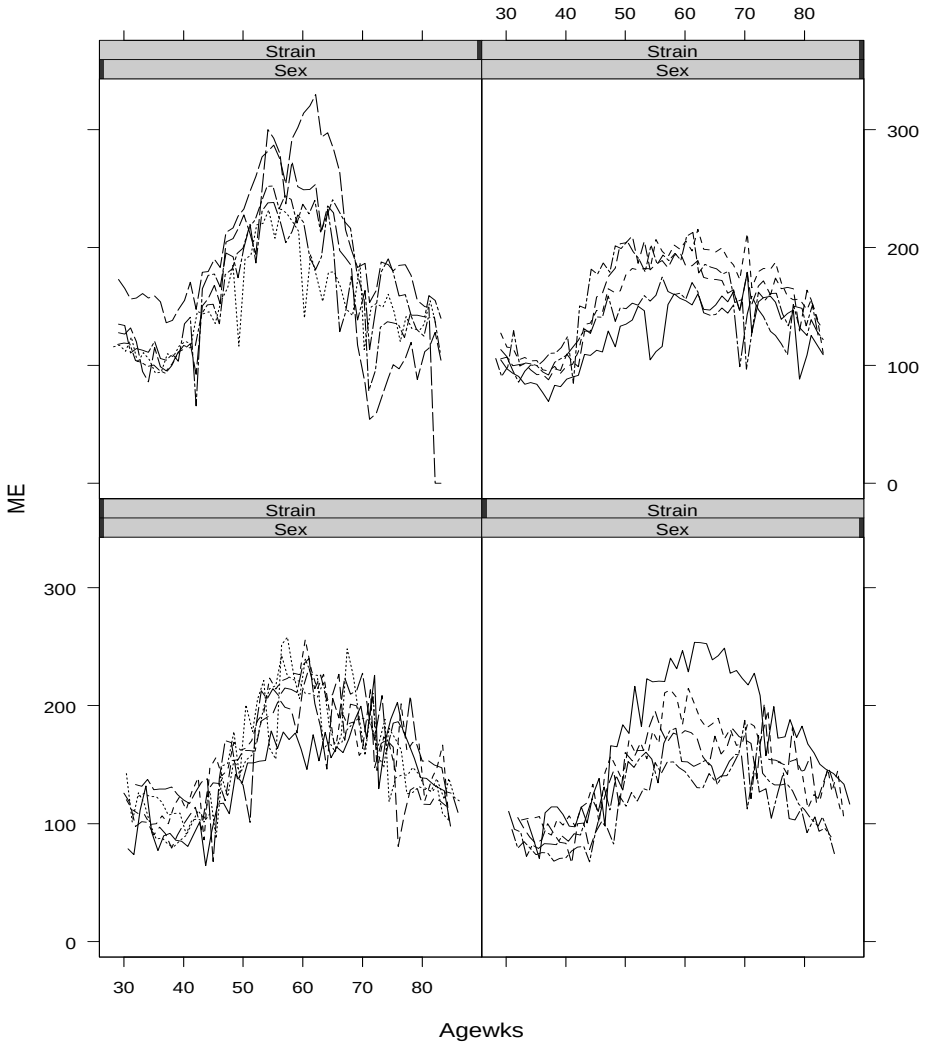
Traces display fanning out suggesting random animal × time effect.

```
ewt ~ weekno|Ca*feed,type='l',data=ewt,
panel=panel.superpose,groups=an)
```



Traces are parallel suggesting random animal effect.

```
xyplot(ME ~ Agewks|Sex*Strain,data=deerdf,type='l',panel=panel.superpose,groups=Tag)
```

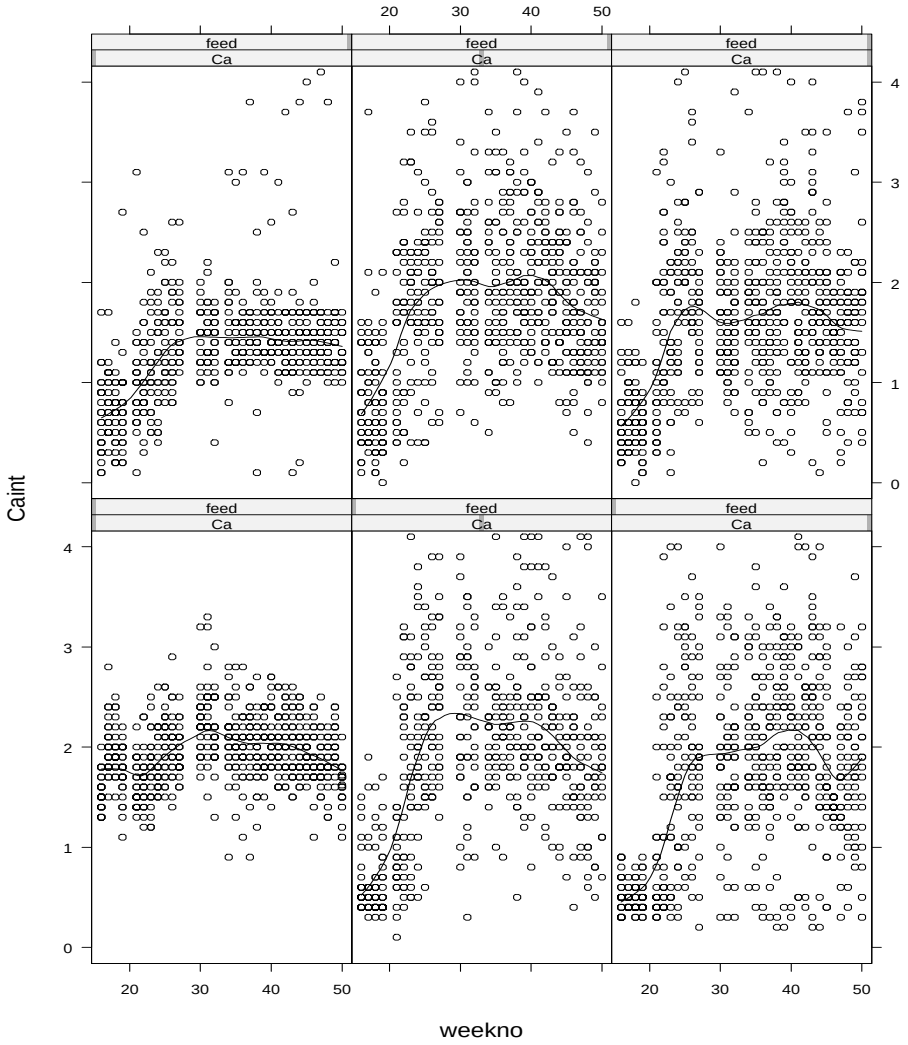


Separation at maximum is bigger than at the start and the traces close up again.

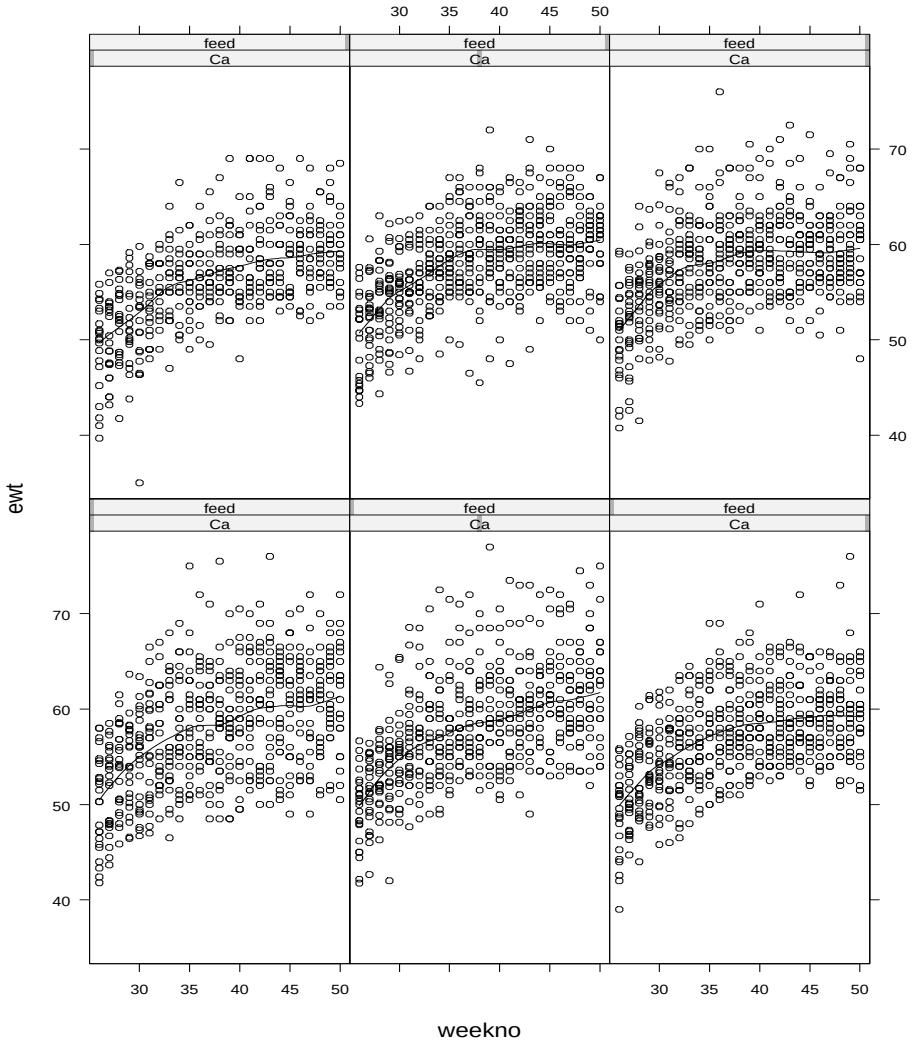
8.3.2 Systematic components

Models for the systematic effect are gleaned from loess curves which in the following cases suggest splines to represent the effect over time.

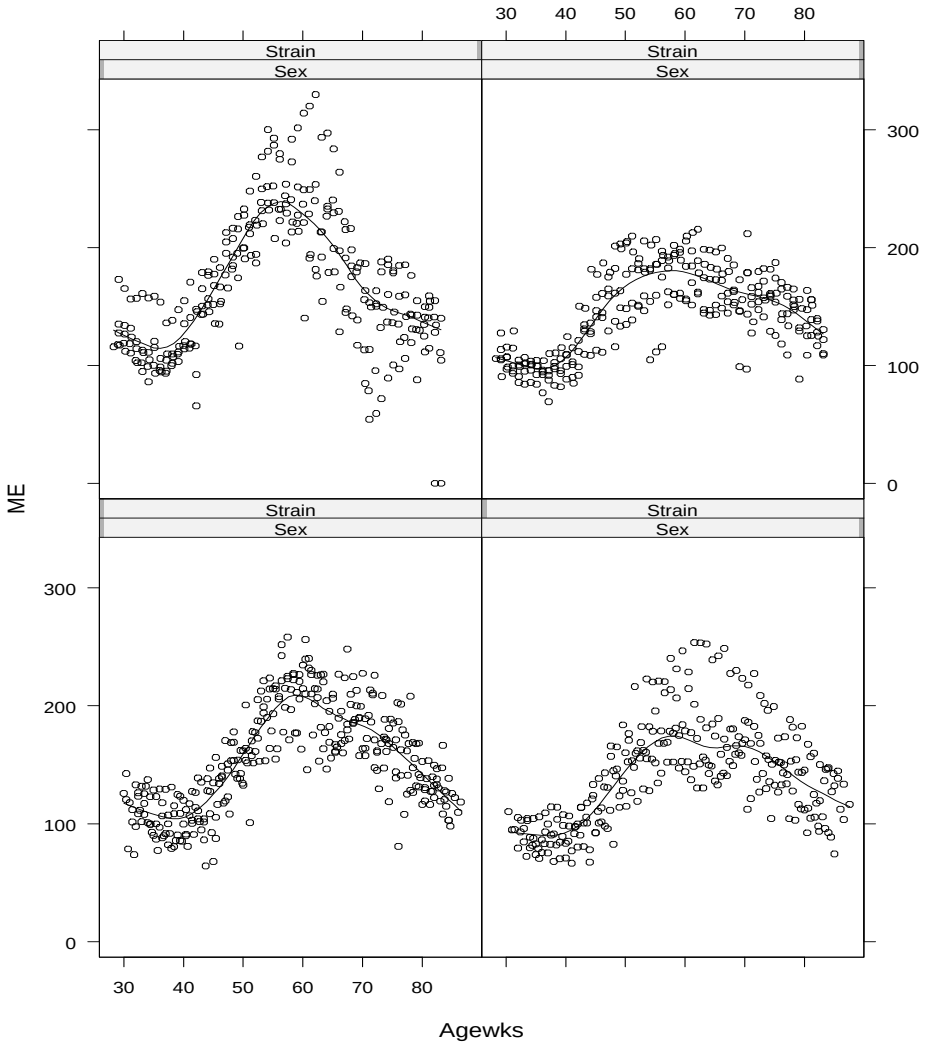
```
xyplot(Caint ~ weekno|Ca*feed,data=Caint,ylim=c(0,4),
panel=function(x,y){panel.xyplot(x,y);panel.loess(x,y,span=1/4) })
```



```
xyplot(ewt ~ weekno|Ca*feed,data=ewt,
panel=function(x,y){panel.xyplot(x,y);panel.loess(x,y,span=1/4) })
```



```
xyplot(ME ~ Agewks|Sex*Strain,data=deerdf,
        panel=function(x,y){panel.xyplot(x,y);panel.loess(x,y,span=1/4) })
```



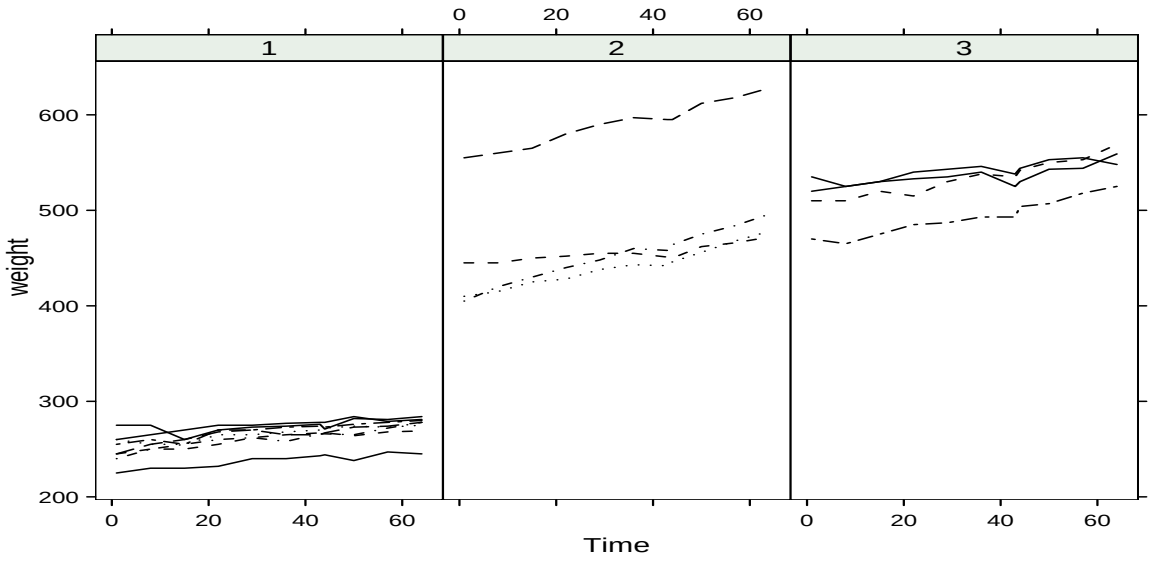
8.4 Linear Mixed Models for repeated Measures

Example 8.1

The example shown in Figure 8.1 are the body weights of rats that were published by Crowder and Hand ([??]) and the data set is included in the nlme library of R , see [7]. The essential R code loads the nlme, grid and lattice libraries and plots the traces of each rat with a separate frame for each Diet.

```
library(nlme);library(grid);library(lattice)
data(BodyWeight)

xyp2 <- xyplot(weight ~ Time|Diet,data=BodyWeight,
 panel=panel.superpose,groups=Rat,type='l',layout=c(3,1))
print(xyp2)
```

Figure 8.1: Profiles of rat body weights over time for 3 treatments



Exploratory plots such as these are necessary to see the structure of both systematic and random effects. We observe that
(i) the response is approximately linear over time for all rats,
(ii) that diets 2 and 3 produce heavier rats than diet 1,
(iii) that growth rate for diet 1 is less that the rates for diets 2 and 3,
(iv) that the variability in diets 2 and 3 is greater than in diet 1,
(v) within Diets slopes are similar,
(vi) differences amongst rats at the beginning are preserved throughout the trial.

We can visualise the mean response as a weighted average of the individual profiles provided that the amongst rats variance should figure in the averaging so that the top rat in diet 2 does not unduly influence the mean response for that diet, for instance.

8.5 A statistical model for repeated measures

Although the rat bodyweights data are simplistic, they suffice to identify the essential ingredients of a model and the steps required to fit the model to the data.

In following a unit through time, the response will be a combination of these effects,

(i) treatment,
(ii) time,
(iii) treatment × time,
(iv) unit,
(v) unit × time

The remaining component is the residuals which may not be independent if an observation at say time t_i is influenced by the outcome at a previous sampling, say $t_i - 1$. This occurs when there is feedback. Figure **??** shows the responses of subjects' elbow flux (blood flow through the elbow) over 15 minutes. Peaks are followed by troughs; the "saw-toothing" is measurement error and successive observations are negatively correlated. If we identify this correlation, the resultant fit will be a smooth track through the observations but otherwise the model will struggle to interpret the saw-toothing as systematic information if residuals are assumed independent.

The linear response for each rat is found conveniently in nlme() by the lmList() function and these individual fits may help in explaining random effects when the model is fitted. Confidence intervals for these estimates are derived using intervals().

```
bwlist1 <- lmList(weight ~ Time,data=BodyWeight)
bwint1 <- intervals(bwlist1)
print(bwlist1)
print(bwint1)
Call:
  Model: weight ~ Time | Rat
   Data: BodyWeight

Coefficients:
   (Intercept)  Time
2          227 0.330
3          247 0.398
4          255 0.330
1          245 0.484
8          252 0.409
5          256 0.406
6          264 0.318
7          268 0.202
11         443 0.363
9          407 1.011
10         408 1.341
12         552 1.148
13         462 0.919
15         524 0.493
14         526 0.315
```

```
16         503 0.905

Degrees of freedom: 176 total; 144 residual
Residual standard error: 4.44
```

We fit the model (**??**) to these data with the lme() function. In the next sequence of code, 2 models are fitted. The purpose of the first is to get the AOV and the second is to get the regression coefficients in a convenient form.

```
mod1 <- lme(weight ~ Diet*Time,data=BodyWeight,random=~ Time|Rat)
mod2 <- lme(weight ~ Diet/Time-1,data=BodyWeight,random=~ Time|Rat)

print(anova(mod1))
print(summary(mod2))
```

```
            numDF denDF F-value p-value
(Intercept)     1   157  1713.2  <.0001
Diet            2    13    85.4  <.0001
Diet:Time       3   157    32.6  <.0001
```

The fixed effects are those represented by the symbol β in equation (**??**).

```
Fixed effects: weight ~ Diet/Time - 1
           Value Std.Error  DF t-value p-value
Diet1      251.7    13.094  13  19.219  <.0001
Diet2      452.3    18.518  13  24.426  <.0001
Diet3      503.7    18.518  13  27.202  <.0001
Diet1:Time   0.4     0.091 158   3.946   1e-04
Diet2:Time   1.0     0.129 158   7.491  <.0001
Diet3:Time   0.7     0.129 158   5.105  <.0001
```

The random effects are saved in `mod1$coefficients$random` and the parameters designated **u** in (**??**).

```
$Rat
   (Intercept)    Time
2      -25.100 -0.0282
3       -4.720  0.0355
4        3.759 -0.0278
1       -6.501  0.1157
8        0.207  0.0456
5        4.287  0.0431
6       12.265 -0.0378
7       15.803 -0.1461
11     -11.043 -0.5608
9      -45.178  0.0403
10     -43.585  0.3475
```

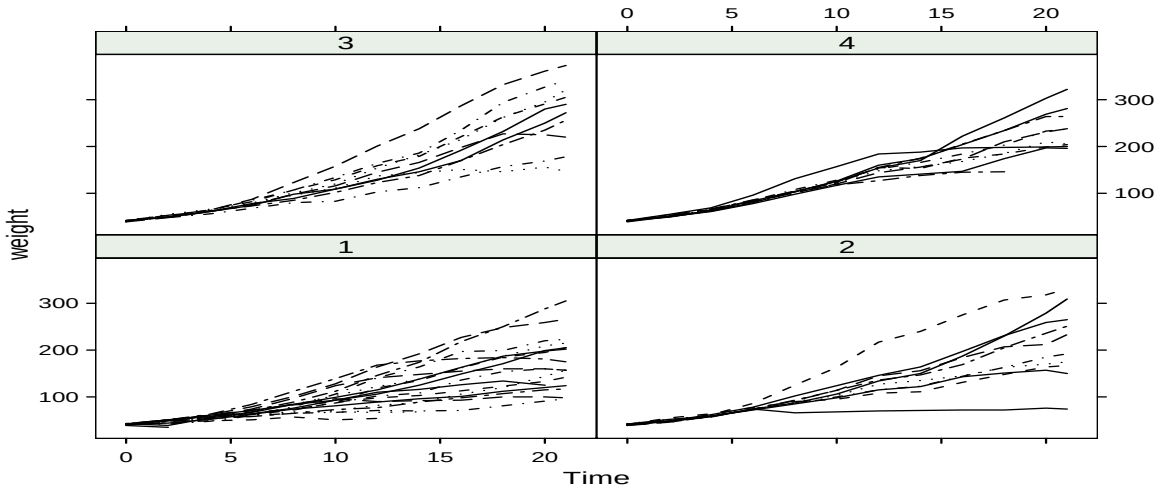
```
12      99.806  0.1730
13     -40.910  0.2406
15      19.478 -0.1524
14      21.131 -0.3182
16       0.300  0.2300
```

Example 8.2

These data are chick weights and are akin to the rats data but with more features. They are also included as part of the nlme package, Pinheiro and Bates [7]. In the following R code, we load the library nlme to get the data and to use the lme() function. Plotting is done using the xyplot() function which is part of the lattice library. Many of these plotting functions also rely on functions from the grid library so we load it as well.

```
library(nlme);library(grid);library(lattice)

data(ChickWeights); print(names(ChickWeight))

xyp2 <- xyplot(weight ~ Time|Diet,data=ChickWeight,
            panel=panel.superpose,groups=Chick,type='l')
print(xyp2)
```

Figure 8.2: Chick weights over time.

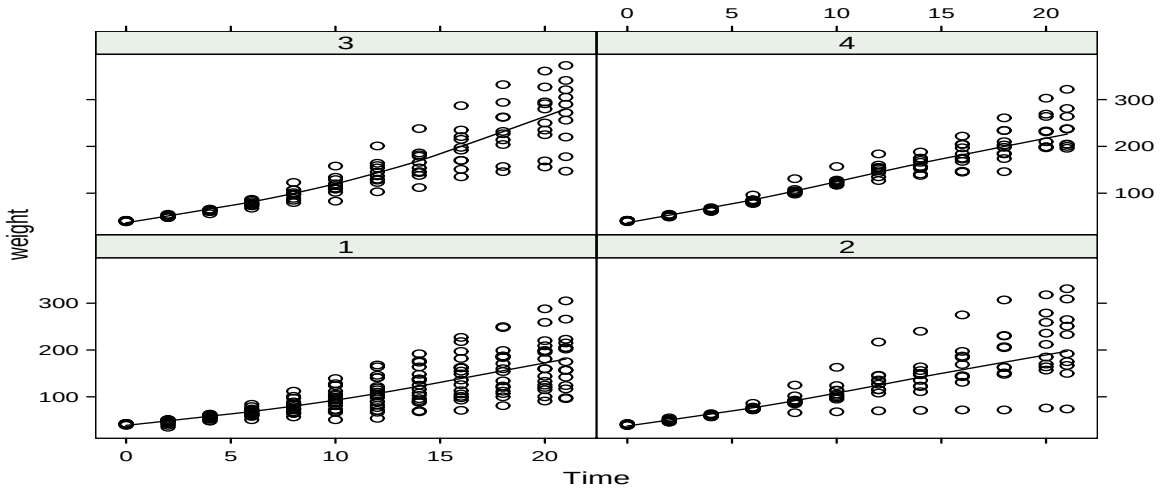


These traces suggest that the systematic effect of Time within each Diet be curved or be piecewise linear with change point at about 12 days. There is a clear Chick × Time interaction as shown by the fanning out of the traces indicating that some chicks grow faster than others.

A clearer view of the systematic effect is obtained when the data are plotted as points and the trend is plotted as a loess curve. This can be achieved by instructing the function of how to draw each plot with the `panel=` option

```
xyp1 <- xyplot(weight ~ Time|Diet,data=ChickWeight,
 panel=function(x,y){panel.xyplot(x,y);panel.loess(x,y)})
print(xyp1)
```

Figure 8.3: Loess plot of trend of chick weights over time.



In Figure 8.2, observe that in Diets 1 and 4 the lightest chick drops out before the trial is complete. That sort of dropout can affect the mean profile because the upward trend may be more due to the remaining chicks that due to the effect of Diet over time. The necessary adjustments for dropouts are not covered in this section but the issue here is that the exploratory data plots revealed the likely bias due to dropouts.

Example 8.3

These data are repeated measurements of Metabolizable energy of 2 Strains of deer × {Male,Female} where the objective was to use the profile to detect the weeks where the body temperature was heightened. The researcher was questioning whether the red deer rutted at a different time to the Pere David deer.

```
xyp1 <- xyplot(ME ~ Agewks|Sex*Strain,data=deerdf,
  panel=panel.superpose,groups=Tag,type='l')
print(xyp1)
```

These plots suggest that the systematic part be represented by a spline curve. Note also that the effects due to animals (called Tag in the data frame) are greater when the curve is maximum than when minimum. The following code has worked but at severe stress to the computer so I am not recommending it for home computers. The bs() function fits a B-spline and the `Tag*bs(Agewks,df=4)` interaction accounts for the changing Tag effects with time.

```
model2 <- lme(ME ~ Strain*Sex*bs(Agewks,df=5),
         random =~ bs(Agewks,df=4) | Tag,data=deerdf)
```

Figure 8.4: Profiles of ME for 2 strains and M,F deer.

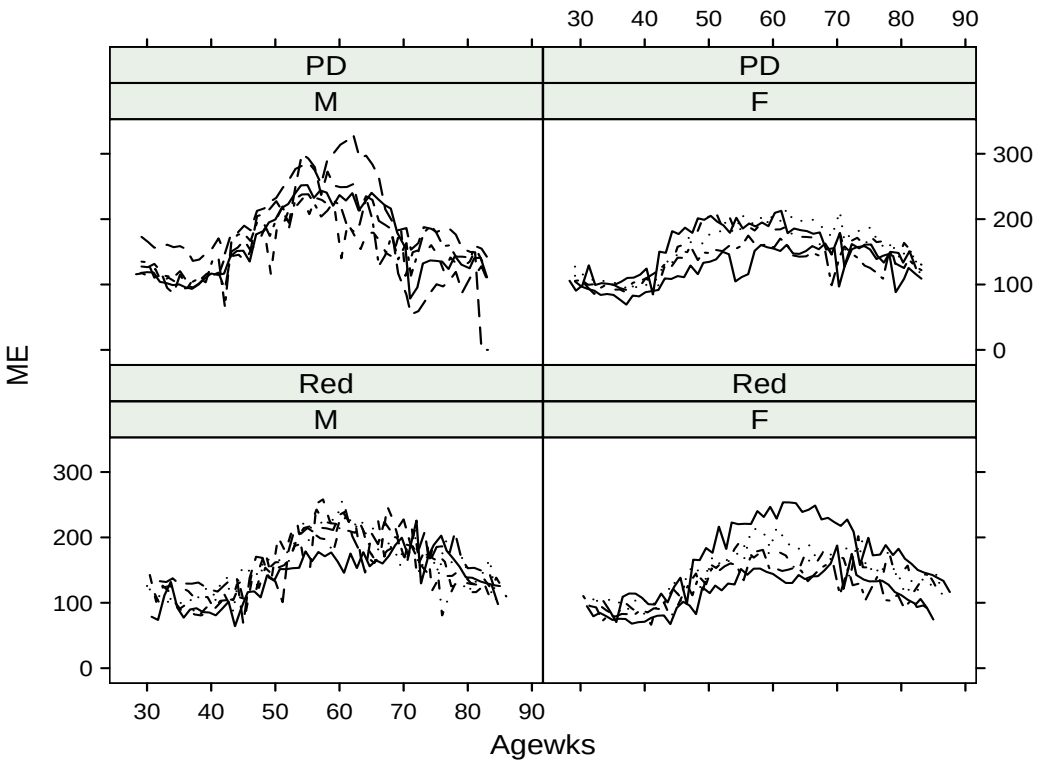
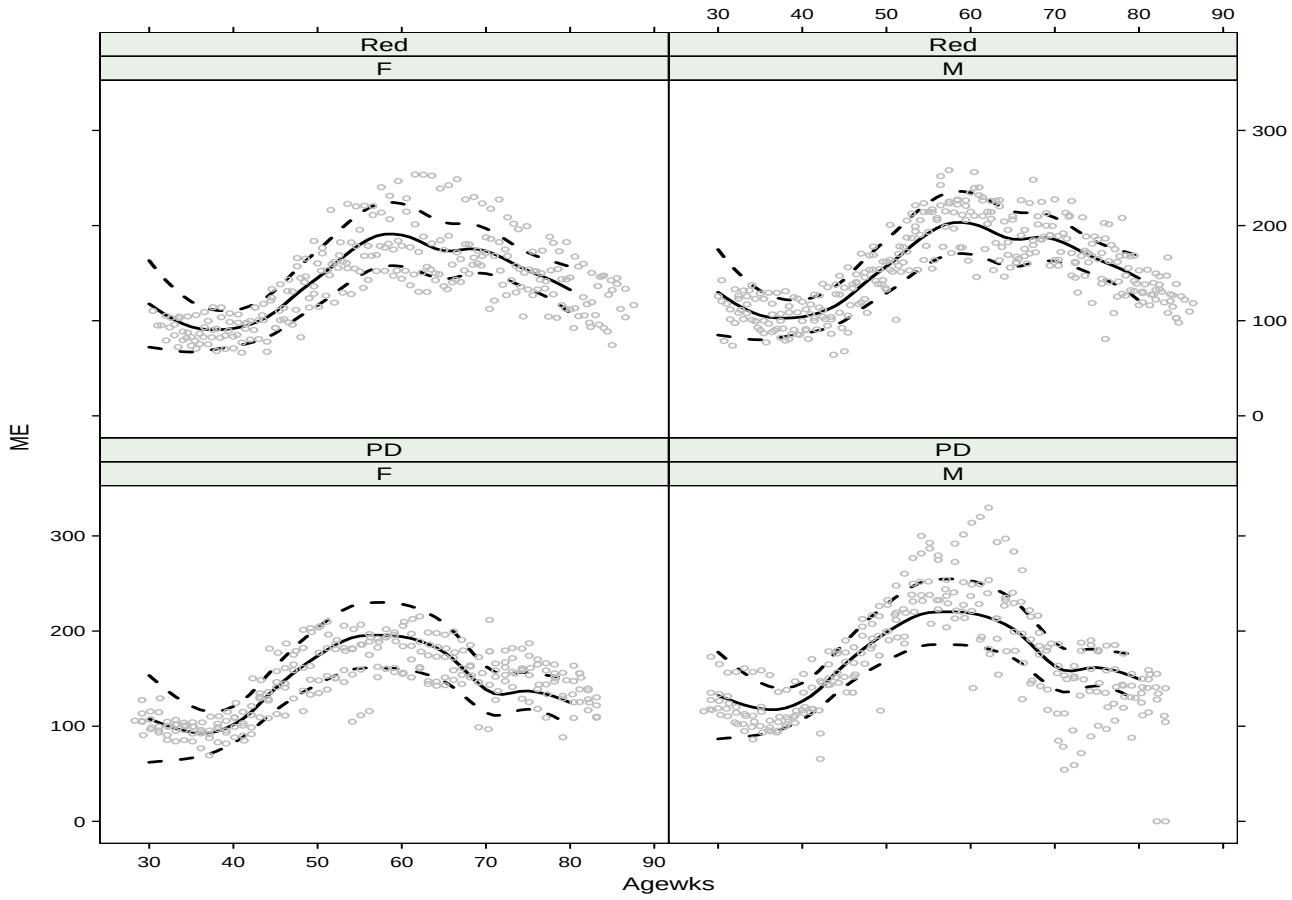


Figure 8.5: Fitted profiles and observed ME for 2 strains and M,F deer.



Chapter 9

Generalised Linear Mixed Models

The content of this is drawn from [2].

The ideas of a linear mixed model extend to regression models for discrete and non-Gaussian continuous data. In the linear model, a random effects model is appropriate if the coefficients can be thought of as a sample from a distribution. There must be repeat samples from each experiment unit to define a random effects model.

In the non-Gaussian setting it is assumed that the data for an experiment unit are independent observations following a GLM but regression coefficients can vary from unit to unit according to a distribution. So the assumptions are:-

1. The conditional distribution of Y_{ij} given u_i follows a distribution from the exponential family with density $f(y_{ij}|\mathbf{u}_i, \boldsymbol{\beta})$

$$f(y_{ij}|\mathbf{u}_i) = \exp\left[\{(y_{ij}\theta_{ij}) - \psi(\theta_{ij}))\}/\phi + c(y_{ij};\phi)\right]$$

$$\begin{aligned} \mu_{ij} &= E(Y_{ij}|\mathbf{u}_i) = \psi'(\theta_{ij}) &&= g^{-1}(x_{ij}\boldsymbol{\beta} + z_{ij}\mathbf{u}_i) && (9.1)\\ v_{ij} &= \mathrm{var}(Y_{ij}|\mathbf{u}_i) = \psi''(\theta_{ij})\phi &&= v(\mu_{ij})\phi && (9.2)\\ & && && (9.3) \end{aligned}$$

   where $g()$ is the link function ($g^{-1}()$ is the inverse of the link function) and $v()$ is the variance function. The vectors $\mathbf{x}$ and $\mathbf{z}$ have dimensions p and q respectively.

2. Given $\mathbf{u}_i$, the repeated measurements $Y_{i_1}, \ldots, Y_{n_i}$ are independent.

3. The $\mathbf{u}_i$ are independent and identically distributed with density function $f(\mathbf{u}_i; \mathbf{G})$. Commonly, $f(\mathbf{u}_i; \mathbf{G})$ is a normal distribution with zero mean and variance matrix $\mathbf{G}(\boldsymbol{\alpha})$.

Correlation amongst observations from a unit, $Y_{i1}, \ldots, Y_{i,n_i}$ arises from their sharing unobservable variables $\mathbf{u}_i$. The random effect model is most useful whenthe objective is to make inference about individuals rather than population averages.

9.1 Inference

Maximum Likelihood

The likelihood of the data expressed as a function of unknown parameters is

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{y}) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\alpha}) d\mathbf{u}_i$$

It is the integral over the unobserved random effects of the joint distribution of the data and random effects. With Gaussian data, the integral has a closed form solution and relatively simple methods exist for maximising the likelihood or restricted likelihood. With non-linear models, numerical techniques are needed.

We consider the random effects as missing data so that the 'complete' data for a unit is $(\mathbf{y}_i, \mathbf{U}_i)$. Denote $\ell = \log(L)$ and

$$\mu_{ij}(\mathbf{U}_i) = E(y_{ij}|\mathbf{U}_i) = g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{U}_i)$$

The score equation for $\boldsymbol{\beta}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}|\mathbf{y}\mathbf{U}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij}\{y_{ij} - \mu_{ij}(\mathbf{U}_i)\} = 0$$

The score equation for $\mathbf{G}$ is

$$S_{\mathbf{G}}(\boldsymbol{\beta}, \boldsymbol{\alpha}|\mathbf{y}) = \frac{1}{2}\mathbf{G}^{-1} \left\{ \sum_{i=1}^{m} E(\mathbf{U}_i\mathbf{U}_i'|\mathbf{y}_i) \right\} \mathbf{G}^{-1} - \frac{m}{2}\mathbf{G}^{-1} \tag{9.4}$$

These are solved using the E-M algorithm. In the estimation step, the expectations are evaluated using current parameter values and this may involve multivariable integration of large dimension. This will usually be done by Monte-Carlo integration.

Quasi-likelihood

An alternative strategy that avoids the problem of integration is to use conditional models rather than conditional means in the score equation for $\boldsymbol{\beta}$. This is equivalent to approximating $f(\mathbf{y}_i|\mathbf{U}_i)$ by a normal distribution with with the same mode and curvature.

Let

$$\begin{aligned} v_{ij} &= \text{var}(y_{ij}|\mathbf{U}_i) \\ Q_i &= \text{diag}\{v_{ij} g'(\mu_{ij})^2\} \\ \zeta_{ij} &= g(\mu_{ij}) + (y_{ij} - \mu_{ij}) g'(\mu_{ij}) \quad j = 1, \ldots, n_i \end{aligned}$$

$$\mathbf{V}_i = Q_i + bZ_i \mathbf{G} \mathbf{Z}_i'$$

The matrix $\mathbf{V}_i$ is a $n_i \times n_i$ matrix and $\mathbf{Z}_i$ is a $n_i \times q$ matrix whose jth row is $\mathbf{z}_{ij}$.

For a fixed $\mathbf{G}$, updated values of $\boldsymbol{\beta}$ and $\mathbf{U}$ are obtained by iteratively solving

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_1^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i' \mathbf{V}_i^{-1} \boldsymbol{\zeta}_i \tag{9.5}$$

$$\hat{\mathbf{U}}_i = \mathbf{G}\mathbf{Z}_i \mathbf{V}_i^{-1} (\boldsymbol{\zeta}_i - \mathbf{X}_i \boldsymbol{\beta}) \tag{9.6}$$

These are the mixed model equations applied to the transformed variable $\boldsymbol{\zeta}$.
From equation (9.4),

$$\hat{G} = m^{-1} \sum_{i=1}^{m} E(\mathbf{U}_i \mathbf{U}_i' | \mathbf{y}_i) \tag{9.7}$$

$$= m^{-1} \sum_{i=1}^{m} E(\mathbf{U}_i|\mathbf{y}_i) E(\mathbf{U}_i|\mathbf{y}_i)' + m^{-1} \sum_{i=1}^{m} \mathrm{var}(\mathbf{U}_i|\mathbf{y}_i) \tag{9.8}$$

Plugin values are

$$E(\mathbf{U}_i|\mathbf{y}_i) = \hat{\mathbf{U}}_i$$

$$\mathrm{var}(\mathbf{U}_i|\mathbf{y}_i) = m^{-1} \sum_{i=1}^{m} (\mathbf{Z}_i' Q_i^{-1} \mathbf{Z}_i + G^{-1})^{-1}$$

The approximation gives reasonable estimates of $\boldsymbol{\beta}$ but the approximation is not reliable if there are few observations and the density of the transformed random variable is far from normal.

9.2 Penalised Quasi-likelihood methods

An implementation of the above technique is the `glmmPQL` function in the `MASS` library. this is a wrapper that implements linearization within the `lme()` function.

The data in the file `AlmondFlowers.txt` are the frequencies of flower stages from 162 trees on August 24th, 2006. The data were determined from photographs and the purpose was to select late-flowering trees that would minimise frost damage. The first 10 trees are listed in Table 9.1

The flower stages are ordered categories which can be analysed as a proportional odds logistic regression. This requires that we calculate the cumulative counts across the ordered categories. The trees are random samples so enter the model as a random effect.

$$\mathrm{logit}(p) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U}$$

where p are the cumulative probabilities, $\mathbf{X}$ is the design matrix for categories and $\boldsymbol{\beta}$ are the category effects, $\mathbf{Z}$ is the indicator matrix for trees and $\mathbf{U}$ are the random tree effects.

The following R program reads the data, transforms to cumulative counts, reshapes the data into the long format prior to doing the analysis using `glmmPQL`.

Table 9.1: Frequencies of Flower stages at 1 sampling

No	TreeTag	Date	Bud	PtlTip	HalfOpen	Anth	X1stPtlFall	PostAnth	OvarySwell	Total
1	1at02	1	0	0	2	84	18	29	0	133
2	1at03	1	0	0	0	45	17	87	0	149
3	1at04	1	3	12	2	67	2	1	0	87
4	1at05	1	4	11	0	118	11	9	0	153
5	1at09	1	68	53	25	33	1	0	0	180
6	1at12	1	49	49	20	43	0	0	0	161
7	1at16	1	1	1	1	80	8	1	0	92
8	1at18	1	24	55	32	161	0	0	0	272
9	1at19	1	52	57	62	117	0	0	0	288
10	1at20	1	1	1	2	82	15	40	0	141

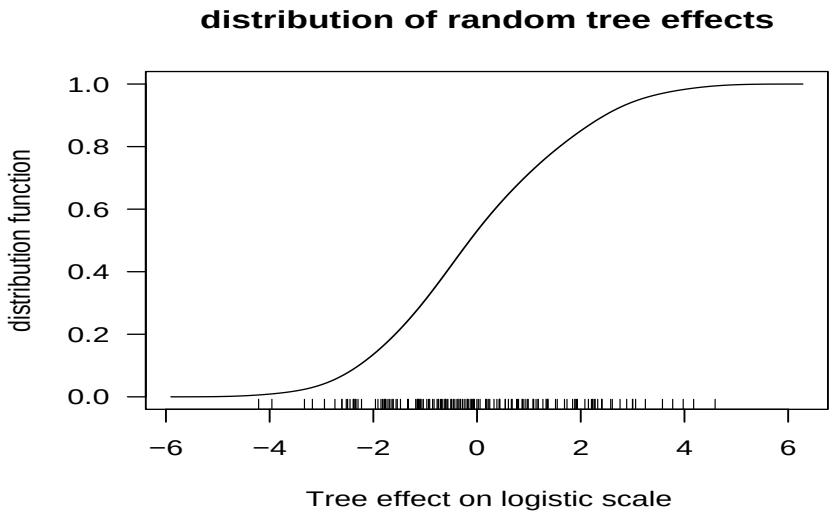
```
#_____________ AlmondFlowers.R _________________
options(digits=2)
library(MASS);library(nlme)

FlT <- read.table("AlmondFlowers.txt",header=T)
vnames <- names(FlT)
FlT[,4:10] <- t(apply(FlT[,4:10],MAR=1,FUN=cumsum)  )

VFlT <- reshape(FlT,varying=list(vnames[3:9]),v.name="Counts",direction="long",timevar="Categor
 times=1:7)
VFlT$Category <- factor(VFlT$Category,labels=vnames[3:9])

oT <- order(VFlT$TreeTag)
VFlT <- VFlT[oT,]

Mod1 <- glmmPQL(Counts/Total ~ Category,data=VFlT,random=~ 1|TreeTag,family="binomial",weights=
```

The next few lines of code extract the fixed and random effects and plots the distribution function.

```
theta <- fixef(Mod1)
alpha <- ranef(Mod1)
oalpha <- order(unlist(alpha))
Tree.effects <- data.frame(Tree=(row.names(alpha))[oalpha],effect=alpha[oalpha,])
print(Tree.effects)
tree.density <- density(Tree.effects$effect)
tree.distribution <- tree.density
tree.distribution$y <- cumsum(tree.density$y)
tree.distribution$y <- tree.distribution$y/max(tree.distribution$y)
```

```
 plot(tree.distribution$y~tree.distribution$x,main="distribution of random tree effects",
         type='l',xlab="Tree effect on logistic scale",las=1,ylab="distribution function")
rug(side=1,x=Tree.effects$effect)
```

Figure 9.1: Distribution of tree effects for flower stages



Negative tree effects are associated with late flowering. From the model we observe that when logit(p) is reduced, this is a shift to the left of the distribution function and a shift to the left is to the early stages of flowering. That is on August 26th, those trees found to have negative tree effects will not be highly represented in the later flowering stages; they have not developed too early.

(The final analysis of these data used the `bayespolr` function in `arm` but the model given here serves as an example.)

9.3 GEE

In cases where predictions are for population averages rather than for individuals, a marginal model can be fitted using a GEE (Generalized Estimating Equation). [2], [5]

In a mixed model, correlation was represented by the shared random effect across repeated measurements. For a GEE, the correlation structure has to be supplied. Suppose that there are n repeated measurements and that the correlations amongst these are represented by a $n \times n$ matrix, $R(\boldsymbol{\alpha})$. The correlation parameters $\boldsymbol{\alpha}$ are in an $s \times 1$ vector and fully characterise $R(\boldsymbol{\alpha})$.

The variance function is $\psi''(\theta_{it})/\phi$ (see (9.2)). The *unscaled* (i.e. omit the scale parameter ϕ) matrix of variance functions across the repeated measures, $t = 1 \ldots, n_i$, from unit i is $A_i = \text{diag}(\psi''(\theta_{it}))$.

Define

$$V_i = A_i^{\frac{1}{2}} R(\boldsymbol{\alpha}) A_i^{\frac{1}{2}}/\phi$$

which is the covariance of $\mathbf{Y}_i$ if $R(\boldsymbol{\alpha})$ is true.

The estimating equations for a GLM are of the form,

$$\sum_{i=1}^{m} \mathbf{X}_i^T \Delta_i (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

where

$$\Delta_i = \frac{d\{\psi'(\theta_{it})\}}{d\eta_{it}}$$

and η_{it} is the linear predictor.

This is extended to contain the correlations to form the GEE,

$$\sum i = 1^m (A_i \Delta_i \mathbf{X}_i)^T V^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{9.9}$$

Equation (9.9) can be expressed as a function of β alone if $\boldsymbol{\alpha}$ is replaced by $\hat{\boldsymbol{\alpha}}$ which is $m^{-\frac{1}{2}}$ consistent.[1]. Also ϕ is replaced by a $m^{-\frac{1}{2}}$ consistent estimator $\hat{\phi}$ when β is known.

The estimating algorithm is evaluated using modified Fisher scoring.

1. Start with estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\phi}$ and $\boldsymbol{\beta}_0$.
2. Estimate a new $\boldsymbol{\beta}$, conditional on the previous estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, $\hat{\phi}$.
3. New estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$, conditional on the current estimate $\hat{\boldsymbol{\beta}}$.
4. Iterate

The updated estimate of $\boldsymbol{\beta}$ is estimated using modified Fisher scoring.

$$\hat{\boldsymbol{\beta}}_{j+1} = \hat{\boldsymbol{\beta}}_j - \left\{ \sum_{i=1}^{m} (A_i \Delta_i \mathbf{X}_i)' \tilde{V}_i^{-1} (A_i \Delta_i \mathbf{X}_i) \right\}^{-1} \left\{ \sum_{i=1}^{m} (A_i \Delta_i \mathbf{X}_i)' \tilde{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right\} \tag{9.10}$$

where $\tilde{V}_i(\boldsymbol{\beta}) = V_i[\beta, \hat{\boldsymbol{\alpha}}\{\beta, \hat{\phi}(\boldsymbol{\beta})\}]$ The matrices in equation (9.10) are functions of $\boldsymbol{\beta}$.

[1] $m^{\frac{1}{2}}(\hat{\alpha} - \alpha) = \mathcal{O}_p(1)$

Estimators of α and ϕ.

Pearson residuals are given by

$$\hat{r}_{it} = \{y_{it} - \psi'(\hat{\theta}_{it})\}/\{\psi''(\hat{\theta}_{it})\}^{\frac{1}{2}}$$

where $\hat{\theta}_{it}$ depends upon the current value for β. The scale parameter ϕ is estimated by

$$\hat{\phi}^{-1} = \sum_{i=1}^{m}\sum_{t=1}^{n_i} \hat{r}_{it}^2/(N-p) \quad (N = \sum_i n_i, p = \text{regression df}).$$

An empirical estimate of R is

$$\hat{R}_{uv} = \sum_{i=1}^{m} \hat{r}_{iu}\hat{r}_{iv}/(N-p)$$

Suppose that we wish to model the correlation structure as a function of the parameters α and the distance between observations,

$$\text{corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$$

.K-Y Liang Since $E(\hat{r}_{it}\hat{r}_{it'}) \approx \alpha^{|t-t'|}$, α can be estimated as the slope in the regression of $\log(\hat{r}_{it}\hat{r}_{it'})$ on $\log|t-t'|$.

Hence we see the algorithm iterates between estimating $(\boldsymbol{\beta}|\hat{\boldsymbol{\alpha}}, \hat{\phi})$ and $(phi, \boldsymbol{\alpha}|\hat{\boldsymbol{\beta}})$

Example

The data in `PastureSpecies.txt` contains percentages of 6 groups of pastures measured at 7 times from 3 farming systems (A, B, C). Within each farm type, there are several paddocks with a Farm × Paddock being an experimental unit.

These are analysed by a binomial GEE using `geeglm` in the `geepack` package.

Table 9.2: First 12 rows of `PastureSpecies.txt`

	Farm	Padd	Days	Group	Frequency	ExpUnit	Total	fDays
1	B	1	0	SowFertResp	54	B.1	100	0
2	B	1	0	NatCoolPeren	4	B.1	100	0
3	B	1	0	NatWarmPeren	0	B.1	100	0
4	B	1	0	YrlGrnPer	26	B.1	100	0
5	B	1	0	WarmAnn	1	B.1	100	0
6	B	1	0	CoolAnn	9	B.1	100	0
7	B	1	256	SowFertResp	42	B.1	100	256
8	B	1	256	NatCoolPeren	21	B.1	100	256
9	B	1	256	NatWarmPeren	3	B.1	100	256
10	B	1	256	YrlGrnPer	23	B.1	100	256
11	B	1	256	WarmAnn	0	B.1	100	256
12	B	1	256	CoolAnn	8	B.1	100	256

The analysis accounts for 2 sources of correlation,

- correlation amongst groups and
- correlation amongst repeated measures

Define

- $R_g(\alpha)$ as the amongst groups correlation matrix and
- R_t as the correlation due to repeated measures from the same experiment unit.

These are estimated separately and the working correlation is $R_B = R_t \otimes R_g$.
The program below does the job in 3 stages:-

1. Run a GLM on the univariate frequency of each group at each sampling and save the residuals. For each sampling, the residuals are saved in a matrix with 6 columns (1 for each group) which is used to calculate the correlation matrix. There are 7 such correlation matrices, one for each sampling, and in this exercise the average of the 7 is used as the estimate of R_g. This may not be a suitable choice as the correlation could change with time. This working correlation matrix is named `workcorr1`.

2. Using `workcorr1`, a GEE analysis is done of the multivariate frequencies from the 6 groups at each sampling. residuals are saved and used to construct the working correlation for amongst samplings, R_t which is named `workcorr2` in the program.

3. The big working correlation matrix, R_B, is calculated as above. (`WORKCORR <-workCorr1 %x% workCorr2`). The GEE analysis of the full data set gives regression coefficients which are the log-odds of pasture presence in one farm system compared to another over time.

Table 9.3: R program for a GEE analysis

```
 library(geepack)
# groups of pasture
Grp.names <- c("SowFertResp","NatCoolPeren","NatWarmPeren","YrlGrnPer","WarmAnn","CoolAnn")
grp.nos <- 1:6
samplings <- c("Mar 00","Dec 00","Dec 01","Dec 02","Feb 03","Feb 04","Feb 05")

PS.df <- read.table("PastureSpecies.txt",header=T)
PS.df$ExpUnit <- interaction(PS.df$Farm,PS.df$Padd)
PS.df$Total <- rep(100,dim(PS.df)[1])
PS.df$fDays <- factor(PS.df$Days)
lDays <- levels(PS.df$fDays);    nDays <- length(lDays)

#   Building a working correlation matrix  Rg %*% Rt
               CORR <- Resids <- Fitted <-  list()
# Do samplings individually
   for (m in 1:nDays){   #  selecting each sampling time
   Sm <- subset(PS.df,subset=PS.df$fDays==lDays[m])   # Sample=m
#  Do each component at a sampling individually
         dim0 <- dim(Sm)[1]
      eachTRmat <- matrix(0,dim0,6,dimnames=list(NULL,Grp.names)  )   # seting up R_g
#########   STAGE 1 ###################
    for (k in 1:6){   #   selecting each pasture group
    SmTk <- subset(Sm, subset=Sm$Group==Grp.names[k] )   # Sample=m, Time=k
# naive GLM for each group at each time
   comp.glm <- glm(Frequency/Total  ~ Farm,data=SmTk,family="binomial",weights=Total)
   eachTRmat[,k] <- residuals(comp.glm,type="pearson")
                  }    # ---------------   end of the k loop
   CorMat <- cor(eachTRmat)
   CORR[m] <- list(CorMat)
#  Averaging the amongst species correlation over times (very rough)
   if(m==1) workCorr1 <- CORR[[1]]/7
   else workCorr1 <- workCorr1 + CORR[[m]]/7
#########   STAGE 2 ######################
#  Now get sets of residuals from each sampling, using the groups as a multivariate
   botanal.model <- geeglm(Frequency/Total ~ Farm/Group,data=Sm,
          family=binomial,id=ExpUnit,zcor=workCorr1,na.action=na.omit)
#  Save Residuals from time-by-time analyses in a list
   Resids[m] <- list(botanal.model$residuals)
                             }  # __________________  end of the m loop
#  the big model with groups and samplings combined
R <- matrix(unlist(Resids),150,7,dimnames=list(NULL,samplings) )
vR <- var(R)
Tcor <- cor(R)
Lower <- Tcor[lower.tri(Tcor)]
workCorr2 <- diag(7)
workCorr2[workCorr2==0] <- mean(Lower)
WORKCORR <-workCorr1 %x% workCorr2
############   STAGE 3 ###################
botanal.bigmodel <- geeglm(Frequency/Total ~ fDays/Group/Farm,data=PS.df,
 family=binomial,id=ExpUnit,zcor=WORKCORR,na.action=na.omit)

summ <- summary(botanal.bigmodel)
beta <- round(summ$coefficients,2)
```

Bibliography

[1] D.R. Cox and N. Reid. Parameter orthogonality and approximate conditional inference. *J. Royal Statist. Soc, B*, 49:1–39, 1987.

[2] P. Diggle, K-Y Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

[3] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchial Models*. Cambridge University Press, 2007.

[4] A. Gilmour, B. Gogel, B. Cullis, S. Welham, and R. Thompson. Asreml user guide. Technical report, VSN International Ltd, Hemel Hempstead, UK, 2002.

[5] K-Y Liang and S Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[6] H.D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545, 1971.

[7] J. Pinheiro and D. Bates. *Mixed Effects Models in S and S-PLUS*. Springer, NY, 2000.

[8] G. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6:15–51, 1991.

[9] David Ruppert, Matt Wand, and Ray Carroll. *Semi-parametric Regression*. Cambridge University Press, 2004.

[10] R. Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78:719–727, 1991.

[11] T. Speed. Comment on robinson (1991). *Statistical Science*, 6:43, 1991.

[12] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, 2004.

[13] A. Verbyla. A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics*, 32:227–230, 1990.

Chapter 10

Appendix - Linear least squares

10.1 Least squares in matrix form

The linear regression relationship between a dependent variable $\mathbf{Y}$ and a set of independent (explanatory) variables $\mathbf{x}_j,\ j = 1 \ldots p$ can be expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \quad , \epsilon \sim N(0, \sigma^2 I) \tag{10.1}$$

where Y is a vector of length n, X is a matrix of dimensions $n \times p$ and β is a vector of length p. For example, for a straight line $E(y_i) = \beta_1 + \beta_2 x_i$ where E represents expected value. In matrix form this becomes

$$\begin{pmatrix} E(y_1) \\ \vdots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

The residual sum of squares (SS) is given by

$$S_r = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{10.2}$$

The least squares estimate of β (the vector of model parameters) is that which gives the least (minimum) value of S_r. Following the standard procedure for minimising functions, differentiating S_r with respect to β gives

$$\frac{\partial S_r}{\partial \boldsymbol{\beta}} \propto \boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Setting this derivative equal to zero gives

$$0 = \boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

$$\text{therefore} \tag{10.3}$$

$$\boldsymbol{X}^T \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{Y} \text{and} \tag{10.4}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{10.5}$$

Equations (10.4) are known as the **normal** equations. These equations assume that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ can be inverted (i.e. is non singular).

Defining the variance/covariance matrix of $\boldsymbol{Y}$ as Var($\boldsymbol{Y}$)

$$\begin{aligned} \text{assume that Var}(\boldsymbol{Y}) &= \sigma^2\boldsymbol{I} \quad \text{where } \boldsymbol{I} \text{ is the identity matrix,} \\ &\quad \text{and } \sigma^2 \text{ is a scalar representing the variance of } y_i \\ \text{and Var}(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T] \ . \end{aligned}$$

[The latter expression assumes that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.]

$$\begin{aligned} \text{Now } \hat{\boldsymbol{\beta}}-\boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\ \text{So Var } (\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T]\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2 \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2. \end{aligned} \tag{10.6}$$

This is the estimate of the variance/covariance matrix of $\hat{\beta}$ i.e. the variance/covariance matrix of the parameter estimates.

10.2 Gauss Markov Theorem

If a linear function $\mathbf{l}^T\boldsymbol{\beta} = \sum_{j=1}^{j=p} l_j\beta_j$ is estimated by a linear combination of observations $\mathbf{m}^T\mathbf{Y} = \sum_{i=1}^{i=n} m_i y_i$, then the unbiased estimator with minimum variance is the corresponding function of the least squares estimates $\mathbf{l}^T\hat{\boldsymbol{\beta}}$.

10.3 Partitioning of parameter vector β

In any vector of parameters $\boldsymbol{\beta}$ it is likely that interest will be centred on some of the parameters. In such circumstances rather than looking to see whether there are differences between all of the parameters one might only be interested in testing for differences between the parameters of interest. One way of approaching this problem is to break down (or partition) the parameter vector into parts. The vector β of equation 10.1 can be split (or partitioned) into two parts

$$\left(\frac{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}\right), \text{where } \boldsymbol{\beta} \text{ is } p\times 1, \text{ and } \boldsymbol{\beta}_1 \text{ is } q\times 1, \ \boldsymbol{\beta}_2 \text{ is } (p-q)\times 1$$

Interest is in the parameter vector $\boldsymbol{\beta}_2$. $\boldsymbol{\beta}_1$ contains only parameters not of interest and which can be eliminated.

$\boldsymbol{X}$ can be partitioned similarly into $(\boldsymbol{X}_1|\boldsymbol{X}_2)$, where $\boldsymbol{X}_1$ is $n\times q$, and $\boldsymbol{X}_2$ is $n\times(p-q)$.

$$\text{Let } \ \boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X} = \left(\begin{array}{c|c} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \hline \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2 \\ \hline \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2 \end{array}\right)$$

$$\text{and } \boldsymbol{C}^{-1} \ = \ \left(\begin{array}{c|c} \boldsymbol{C}^{11} & \boldsymbol{C}^{12} \\ \hline \boldsymbol{C}^{21} & \boldsymbol{C}^{22} \end{array}\right)$$

The normal equations $(\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$ give

$$\left.\begin{array}{rcl} \boldsymbol{C}_{11}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{C}_{12}\hat{\boldsymbol{\beta}}_2 &=& \boldsymbol{X}_1^T\boldsymbol{Y} \\ \boldsymbol{C}_{21}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{C}_{22}\hat{\boldsymbol{\beta}}_2 &=& \boldsymbol{X}_2^T\boldsymbol{Y} \end{array}\right] \quad . \tag{10.7}$$

Fitting β_1 only, ignoring β_2, gives

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{C}_{11}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$$

[Solution of normal equations (10.4) $(\boldsymbol{X}_1^T\boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1^T\boldsymbol{Y}$]

The second equation of (10.7) gives [$\boldsymbol{C}^{22} = (\boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12})^{-1}$ from right of $\boldsymbol{C}^{-1}$]

$$\hat{\boldsymbol{\beta}}_2 = \boldsymbol{C}^{22}(\boldsymbol{X}_2^T - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{X}_1^T)\boldsymbol{Y} \quad .$$

Notice that this expression becomes $\hat{\boldsymbol{\beta}}_2 = \boldsymbol{C}_{22}^{-1}\boldsymbol{X}_2^T\boldsymbol{Y}$ only when $\boldsymbol{C}_{21} = \boldsymbol{0}$ (zero matrix).

$$\text{If } \boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X} = \left(\begin{array}{c|c} \boldsymbol{C}_{11} & 0 \\ \hline 0 & \boldsymbol{C}_{22} \end{array}\right) \tag{10.8}$$

$\boldsymbol{X}_1, \boldsymbol{X}_2$ are orthogonal to each other, i.e. estimation of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are independent of each other.

If $\boldsymbol{X}_1, \boldsymbol{X}_2$ are not orthogonal, estimation of $\hat{\boldsymbol{\beta}}_2$ is related to estimation of $\hat{\boldsymbol{\beta}}_1$.

Fitting $\boldsymbol{\beta}_2$ ignoring $\boldsymbol{\beta}_1$ would give $\hat{\boldsymbol{\beta}}_2 = \boldsymbol{C}_{22}^{-1}\boldsymbol{X}_2^T\boldsymbol{Y}$, fitting $\hat{\boldsymbol{\beta}}_2$ after eliminating $\hat{\boldsymbol{\beta}}_1$ gives $\hat{\boldsymbol{\beta}}_2 = \boldsymbol{C}^{22}(\boldsymbol{X}_2^T - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{X}_1^T)\boldsymbol{Y}$.

$$\begin{array}{rcl} \text{Sum of squares for } \hat{\boldsymbol{\beta}} &=& \boldsymbol{Y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ \text{Sum of squares for } \hat{\boldsymbol{\beta}}_1 &=& \boldsymbol{Y}^T\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 \\ \text{Sum of squares for } \hat{\boldsymbol{\beta}}_2 &=& \hat{\boldsymbol{\beta}}_2^T(\boldsymbol{C}^{22})^{-1}\hat{\boldsymbol{\beta}}_2 \end{array}$$

10.4 Order Notation

This notation refers to orders of magnitudes in asymptotics, i.e. how a statistic, which is a function of sample size, behaves as sample size increases.

$$\begin{aligned} c_n = \mathcal{O}(d_n) &\Rightarrow \lim_{n\to\infty} \frac{c_n}{d_n} \leq \text{Const.} \\ c_n = o(d_n) &\Rightarrow \lim_{n\to\infty} \frac{c_n}{d_n} = 0 \\ c_n \sim d_n &\Rightarrow \lim_{n\to\infty} \frac{c_n}{d_n} = 1 \end{aligned}$$

The limit can also be expressed in this form:-

$$c_n = o(d_n) \Rightarrow \frac{c_n}{d_n} \to 0 \text{ as } n \to \infty$$

For random variables,

$$\begin{aligned} U_n = \mathcal{O}_p(d_n) &\Rightarrow \lim_{n\to\infty} \frac{U_n}{d_n} \text{ is bounded in probability} \\ U_n = o_p(d_n) &\Rightarrow \lim_{n\to\infty} \frac{U_n}{d_n} \to 0 \text{ in probability.} \end{aligned}$$

For the latter, we require that given $\epsilon > 0$, there exists constants k_ϵ and $n_0 = n_0(\epsilon)$ such that if $n > n_0$ $P\{|U_n| < d_n k_\epsilon\} > 1 - \epsilon$.

In particular, $U_n = c + o_p(1) \Rightarrow U_n \xrightarrow{p} 1$.

An important special case is when $\text{var}(U_n) \leq \frac{v}{n}$, for $n \geq n_0$ and for finite $v > 0$. Then $U_n = E(U_n) = \mathcal{O}_p(n^{-\frac{1}{2}})$.

If in addition, $E(U_n) = \mu + \mathcal{O}(n^{-\frac{1}{2}})$, then $U_n = \mu + \mathcal{O}_p(n^{-\frac{1}{2}})$.