Mixed models for metric data

(3402-451)

Prof. Dr. Hans-Peter Piepho

Institute for Crop Science (340)
Biostatistics Unit
Universität Hohenheim
e-mail: piepho@uni-hohenheim.de

Hohenheim, October 2015

A 8014

Mixed models for metric data

Hans-Peter Piepho
(University of Hohenheim, Stuttgart, Germany, piepho@uni-hohenheim.de)

Table of contents **Page no.**

0 Preliminaries 2

1 Motivation of mixed models 4

2 Randomized complete block design (prelude to random blocks) 7
2.1 Linear model 8
2.2 Analysis of variance 9
2.3 Mean comparisons 11

3 Incomplete blocks 13
3.1 The need for adjustment 13
3.2 Adjusted means 17
3.3 Least squares estimation 17
3.4 Recovery of inter-block information and mixed modelling 21
*3.5 Estimation of variance components by ML and REML: a simple example 26
3.6 Intra-block and combined inter-intra-block analysis 32
3.7 Analysis of a non-resolvable augmented design 32
3.8 Analysis of an α-design 36
3.9 Analysis of a resolvable augmented design 38
3.10 Analysis of a resolvable row-column design 39
3.11 Efficiency of resolvable designs 41
3.12 Post blocking 44

4 Sub-sampling 46
4.1 The sorghum experiment 46
4.2 General results for inference on fixed effects and application to examples 50

5 Split-plot design 59
5.1 Randomisation 59
5.2 Basic modelling 60
5.3 Polynomial regression for a split-plot experiment 71

6 Repeated measures 77
6.1 The duckweed experiment 77
6.2 Model selection by likelihood ratio tests and AIC 80
6.3 Inference for fixed effects using selected variance-covariance model 84
6.4 A more complex example with Arabidopsis 87

7 Spatial models for field trials 94
7.1 Revisiting the oats data of Section 3.8 95
7.2 Model diagnosis by semivariogram 96
7.3 The wheat data of Besag and Kempton 98
7.4 Extension in two dimensions 100

8 Random genotype effects 103
8.1 Selection and shrinkage 103
8.2 Best linear unbiased prediction of genetic effects (BLUP) 104
8.3 Genetic covariance in pedigrees 111
8.4 The numerator relationship matrix 112
8.5 Pedigree-based BLUP 113
8.6 BLUP under selection 117
8.7 Multivariate BLUP 118
8.8 GCA and SCA in a diallel 121
8.9 The general BLUP equation and the mixed model equations 123

9 Series of experiments 124
9.1 Basic modelling 124
9.2 Modelling cultivar × location interaction 128

Acknowledgements 133

References 133

A. Exercises for practicals 135

B. A short MIXED manual 156
1. PROC MIXED call 156
2. CLASS statement 156
3. MODEL statement 156
4. RANDOM statement 157
5. REPEATED statement 157
6. LSMEANS statement 157
7. ODS tables 157
8. ODS graphics 158
9. PARMS statement 159
10. Trouble shooting in case of non-convergence 159

C. Solutions to some exercises 163

Preliminaries

During this lecture you will learn how to use the statistical package SAS for fitting mixed models. The package will also be used during the exam. It is therefore necessary that in time you familiarize yourself with the usage of this package. Included in the Appendix of these notes are some technical hints (A) as well as worked examples (C&D). There is also a companion file "Solutions to examples" that has solutions in SAS for the examples used in the lecture. The package has a very extensive help menu which you should consult frequently in case of questions.

In exams, you will be given two problems for the lecture "mixed models" that can and should be solved using the MIXED procedure. Your solution is best written in a MS WORD document, into which you can copy and paste your program code as well as relevant parts of the output. You will also need to explain your solution and give brief interpretations

explaining what you find in the output.

I am expecting that you have had exposure to a first course in statistics plus a second course that deals with experimental design and linear models. A good reference for the background I am expecting is this:

Mead R, Curnow RN, Hasted AM 2002 Statistical methods in agriculture and experimental biology. CRC Press, Boca Raton.

There are many good books that cover material similar to that in this course, e.g.

Schabenberger O, Pierce FJ 2002 Contemporary statistical models in the plant and soil sciences. CRC Press, Boca Raton.

I strongly recommend that you read and consult this this book during the course. There are also some articles posted in ILIAS that you should read.

We are assuming that you have been exposed to a first course in statistics. The Mead et al. book can be used to refresh your memory on basic principles. You may also consult the ebook CAST, which is available on ILIAS or at http://cast.massey.ac.nz/collection_public.html. Please not that the module "Bioinformatics" is NOT a first course in statistics. So if you have not had such a course, or if despite having attended such a course you find that you do not have the necessary background on basic concepts, it is your responsibility to fill any gaps by your own reading and study.

Note that a module at the University of Hohenheim comprises 56 hours "contact time". The workload of a module is 150-180 hours. Thus, for every hour of lecture, you are expected to do about two hours of reading and exercising at your own initiative.

1 Motivation of mixed models

Assume that 100 new wheat cultivars are to be tested in a target region. For this purpose, 15 trial locations are selected from the target region. Assume for simplicity that at each location, trials are laid out in a completely randomized design with 4 replicates. A linear model for yield may be formulated as follows:

$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$ and $y_{ijk} = E(y_{ijk}) + e_{ijk}$, where

y_{ijk} = yield in k-th replicate (k = 1, ..., 4)
at j-th location (j = 1, ..., 15) for the i-th cultivar (i = 1, ..., 100)
μ = general effect (intercept)
α_i = main effect of i-th cultivar
β_j = main effect of j-th location
γ_{ij} = interaction of i-th cultivar and j-th location
e_{ijk} = random deviation associated with y_{ijk}

$E(y_{ijk})$ is the **expected value** of an observation for the i-th cultivar and the j-th location. It is the systematic part of the model. The random deviation e_{ijk}, also sometimes denoted – somewhat pejoratively – as "error", is the random or stochastic component of the model. The error is usually assumed to follow a normal distribution with zero mean and variance σ^2. This may be expressed formally by

$e_{ijk} \sim N(0, \sigma^2)$.

In this linear model the effect e_{ijk} is a random variable, while the other effects (μ, α_i, β_j, γ_{ij}) are fixed effects. The **random effects** (e_{ijk}) can be considered as noise, which is caused by heterogeneity among plots in the experimental field. By contrast, the **fixed effects** (μ, α_i, β_j, γ_{ij}) model systematic differences that are caused by the factors cultivar and location. Under these model assumptions, a classical two-factor ANOVA (analysis of variance) can be performed, in which the method of least squares is used to estimate means (BLUE = best linear unbiased estimation). In such an analysis a central issue is the importance of the interaction location × cultivar. In case of significant interaction, the differences among cultivars are not constant among locations due to environmental and genotypic factors. It is then of interest to find the causes of interaction. The model may then possibly extended to model interaction, e.g., by factorial regression using covariates for the cultivars and for the locations.

In the present example, locations have been randomly sampled, so the factor location may be regarded as random rather than fixed. This perspective may be relevant when one is mainly interested in the mean performance of the cultivars in the target region, while individual locations are of less interest. In this case, locations mainly serve as replicates chosen to be representative of the target. If a factor is random, one usually takes all effects associated with that factor as random. In the case at hand the model then becomes:

$E(y_{ijk}) = \mu + \alpha_i$ and $y_{ijk} = E(y_{ijk}) + \beta_j + \gamma_{ij} + e_{ijk}$

$\beta_j \sim N(0, \sigma^2_\beta)$
$\gamma_{ij} \sim N(0, \sigma^2_\gamma)$

$e_{ijk} \sim N(0, \sigma^2)$

In the analysis according to a model with random environments the comparison of cultivar means is the main focus. Interactions are now random, and they will be part of the error term for the mean comparison. In order to compute this error term, it will first be necessary to estimate the variances of random terms, the so-called variance components. As will be shown, estimation is mostly done by the REML method (restricted maximum likelihood). It is also possible to estimate the random effects by a method known as BLUP (best linear unbiased prediction). Estimation by BLUP yields different results than BLUE for fixed effects.

In the model with fixed cultivars and random environments there are both fixed effects (apart from μ) and random effects (apart from the residual e_{ijk}). For this reason the model is a **mixed model**, as it mixes random and fixed effects. As the example shows, the question as to whether an effect or factor is fixed or random is a central issue in the formulation of a mixed model. Searle et al. (1992, p. 18) give the following decision rules for addressing this issue, including consequences for analysis (Fig. 1):

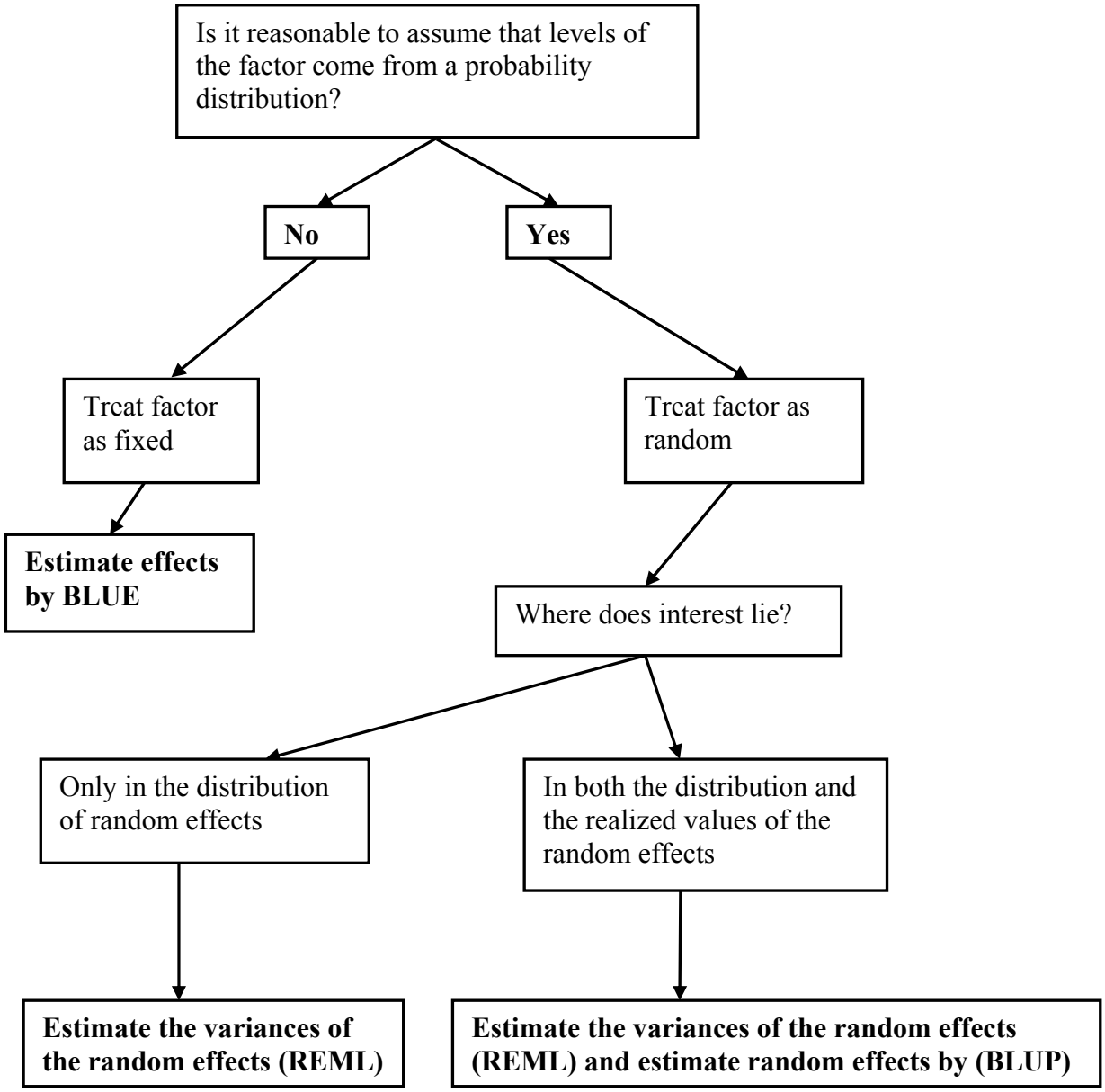


Fig. 1: Decision tree for choice of a factor as fixed or random

Standard analysis can be done by models with fixed effects only, apart from linear error, such as one-way ANOVA, regression, and analysis of covariance. Some examples, where random effects are needed in addition to fixed effects, are as follows.

(1) **Split-plot designs** are designs for factorial experiments, which involve two independent randomization steps. As a result, the linear model needs to have two error terms (main plot and sub-plot error).

(2) **Incomplete blocks designs** can be analysed by a model with random blocks, thus allowing for what is called "recovery of information".

(3) **Series of experiments** are often analysed taking the environmental factor as random thus giving rise to several random effects.

(4) Field measurement often involve **sub-sampling**. For example, several plant samples may be taken from the same plot. Analysis must then account for both sampling and measurement error associated with plants as well as for plot error, giving rise to a mixed model.

Mixed models are also very useful, when **correlation among observations** needs to be modelled. Some examples are as follows.

(5) To exploit **pedigree information** in breeding trials and quantitative genetic studies, it is necessary to regard genotypic effects as random. One can then exploit resemblance among relatives using **genetic correlation** among random genetic effects.

(6) Phenotypic measurements are sometimes taken repeatedly on the same unit (plot, plant etc.). This is called **repeated measurement**. Repeated measurements are typically correlated, and the correlation decays with time.

(7) Analysis of standard experimental designs such as block designs is based on randomization theory. In such analyses, observations within a block are regarded as uncorrelated. While proper randomization always ensures that such an analysis is valid, some improvement may be possible. It is often the case that observations from nearby plots are more similar than observations from plots farther apart. This similarity is due to field trend and may be modelled by a **spatial covariance** or spatial correlation model. Analysis of field trials by spatial methods as embedded in a mixed model framework has gained considerable popularity in recent years, mainly due to the availability of efficient mixed model software.

This text will give an introduction to mixed models, with an emphasis on the use of mixed model software. Some theoretical background will be given as felt essential for an informed use of such software. The development is centered on examples, and the examples cover all of the instances (1) to (7) given above. The examples are mainly related to phenotypic data from experiments with plants. Theoretical background will not be concentrated in a single place for didactical reasons. Instead, theory will be given only as necessary for the example-based development. For ease of reference, sections containing mainly theoretical material are marked with an asterisk in the table of contents.

2 Randomized complete block design

The exposition starts with a fixed effects model for randomized complete block designs. This is needed as a prelude to the analysis of incomplete blocks in Section 3, which requires random blocks. Table 1 shows the layout and result of a block experiment with four genotypes.

Table 1: The field layout and plot yields of a block experiment with four genotypes V_1-V_4 (t/ha) (Clewer and Scarisbrick 2001).

Block 1	V_2	9.8	V_4	9.5	V_3	7.3	V_1	7.4
Block 2	V_3	6.1	V_2	6.8	V_1	6.5	V_4	8.0
Block 3	V_3	6.4	V_2	6.2	V_4	7.4	V_1	5.6

For analysis it is convenient to arrange this dataset as a two-way table with ordered blocks and treatments (genotypes), as shown in Table 2.

Table 2: Yields of block experiment arranged in block × genotype table.

	Genotype			
	1	2	3	4
Block				
1	7.4	9.8	7.3	9.5
2	6.5	6.8	6.1	8.0
3	5.6	6.2	6.4	7.4

In order to assess effects of blocks and genotypes, it is useful to compute marginal means and the overall mean of the table (Table 3).

Table 3: Yields of block experiment arranged in block × genotype table, with marginal means.

	Genotype				
	1	2	3	4	Mean
Block					
1	7.4	9.8	7.3	9.5	8.50
2	6.5	6.8	6.1	8.0	6.85
3	5.6	6.2	6.4	7.4	6.40
Mean	6.5	7.6	6.6	8.3	**7.25** = overall mean

The block means show that block 1 had favorable conditions compared to the other two blocks. Marginal means for the genotypes indicate that genotype 4 has the best mean yield. Obviously, there are both block effects and genotype effects (treatment effects).

2.1 Linear model

The objective of statistical analysis is to dissect the different effects underlying the data. This may be done based on a model that can be written in the form

$$\begin{pmatrix}\text{observed}\\ \text{yield}\end{pmatrix}=\begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix}+\begin{pmatrix}\text{block}\\ \text{effect}\end{pmatrix}+\begin{pmatrix}\text{genotype}\\ \text{effect}\end{pmatrix}+\begin{pmatrix}\text{residual}\\ \text{variation}\end{pmatrix},$$

where block effects and genotype effects are deviations of the particular block or genotype mean from the overall mean.

The effects can be computed by a method called 'sweeping'. By this method, we successively subtract computed effects from the observed data. The process starts by subtracting the overall mean from the observed data in Table 3. The corrected values are displayed in Table 4. Interestingly, when we compute block means of the corrected values in Table 4, we find that the means have an average of zero. The block means of corrected values in Table 4 may be interpreted as block effects. Block 1, which we had found to be a favorable block, has a positive effect (above average), while the other two blocks have negative effects (below average). The mean block effect is zero. Note that the differences of block effects in Table 4 are identical to differences in marginal means for blocks shown in Table 3.

Table 4: Corrected values after subtracting the overall mean in Table 3.

	Genotype				
	1	2	3	4	Mean = block effects
Block					
1	0.15	2.55	0.05	2.25	**+1.25**
2	-0.75	-0.45	-1.15	0.75	**-0.40**
3	-1.65	-1.05	-0.85	0.15	**-0.85**

The sum of squares of the corrected values in Table 4 is the corrected total sum of squares appearing in the analysis of variance to be discussed below. In order to correct for block effects, we subtract block effects (block means) from corrected values in Table 4. The result is shown in Table 5.

Table 5: Corrected values after subtracting the block effects in Table 4.

	Genotype				
	1	2	3	4	
Block					
1	-1.10	1.30	-1.20	1.00	
2	0.35	-0.05	-0.75	1.15	
3	-0.80	-0.20	0.00	1.00	
Mean	**-0.75**	**+0.35**	**-0.65**	**+1.05**	= genotype effects

After the correction for block effects, entries in a row (block) in Table 5 sum to zero. Also,

the genotype means of the corrected values have a zero mean. The genotype means in Table 5 may be interpreted as genotype effects. For example, genotype 4 has a positive effect, corresponding to our earlier finding that it had the highest mean (Table 2). Note that the differences of genotype effects in Table 5 are identical to differences in marginal means for genotypes shown in Table 3.

If we subtract the genotype means (effects) from corrected values in Table 5, we obtain the residuals in Table 6. These residuals are now free of genotype effects, of block effects, and of the overall mean. These effects have been 'sweeped' out. Note that the residuals sum to zero within rows and within columns. Also, the overall sum is zero. The sum of squares of residuals is equal to the residual sum of squares in an analysis of variance to be discussed shortly.

Table 6: Residuals after subtracting the genotype means in Table 5.

	Genotype			
	1	2	3	4
Block				
1	**-0.35**	**+0.95**	**-0.55**	**-0.05**
2	**+0.40**	**-0.40**	**-0.10**	**+0.10**
3	**-0.05**	**-0.55**	**+0.65**	**-0.05**

It is instructive to reassemble the original observations from the various effects and the residuals just computed. These terms are bold-faced in Tables 3 to 6. The 're-assembly' can be written as follows:

$$\begin{pmatrix}\text{observed}\\ \text{yield}\end{pmatrix} = \begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix} + \begin{pmatrix}\text{block}\\ \text{effect}\end{pmatrix} + \begin{pmatrix}\text{genotype}\\ \text{effect}\end{pmatrix} + \begin{pmatrix}\text{residual}\\ \text{variation}\end{pmatrix}$$

$$\begin{pmatrix}7.4 & 9.8 & 7.3 & 9.5\\ 6.5 & 6.8 & 6.1 & 8.0\\ 5.6 & 6.2 & 6.4 & 7.4\end{pmatrix} = \begin{pmatrix}7.25 & 7.25 & 7.25 & 7.25\\ 7.25 & 7.25 & 7.25 & 7.25\\ 7.25 & 7.25 & 7.25 & 7.25\end{pmatrix} + \begin{pmatrix}+1.25 & +1.25 & +1.25 & +1.25\\ -0.40 & -0.40 & -0.40 & -0.40\\ -0.85 & -0.85 & -0.85 & -0.85\end{pmatrix} + \begin{pmatrix}-0.75 & +0.35 & -0.65 & +1.05\\ -0.75 & +0.35 & -0.65 & +1.05\\ -0.75 & +0.35 & -0.65 & +1.05\end{pmatrix} + \begin{pmatrix}-0.35 & +0.95 & -0.55 & -0.05\\ +0.40 & -0.40 & -0.10 & +0.10\\ -0.05 & -0.55 & +0.65 & -0.05\end{pmatrix}$$

More formally, our model can be written as

$$y_{ij} = \mu + b_j + \tau_i + e_{ij} \quad , \qquad (1)$$

where y_{ij} is the observed yield of the i-th genotype in the j-th block, μ is the overall mean, b_j is the effect of the j-th block, τ_i is the effect of the i-th genotype, and e_{ij} is the residual effect corresponding to y_{ij}.

2.2 Analysis of variance

In order to partition the total variation, one may inspect the deviation of observations from the overall mean. Subtracting the overall mean on both sides of our 're-assembly' we may write

$$\begin{pmatrix}\text{observed}\\ \text{yield}\end{pmatrix}-\begin{pmatrix}\text{overall}\\ \text{mean}\end{pmatrix} = \begin{pmatrix}\text{block}\\ \text{effect}\end{pmatrix} + \begin{pmatrix}\text{genotype}\\ \text{effect}\end{pmatrix} + \begin{pmatrix}\text{residual}\\ \text{variation}\end{pmatrix}$$

$$\begin{pmatrix}+0.15 & +2.55 & +0.05 & +2.25\\ -0.75 & -0.45 & -1.15 & +0.75\\ -1.65 & -1.05 & -0.85 & +0.15\end{pmatrix} = \begin{pmatrix}+1.25 & +1.25 & +1.25 & +1.25\\ -0.40 & -0.40 & -0.40 & -0.40\\ -0.85 & -0.85 & -0.85 & -0.85\end{pmatrix} + \begin{pmatrix}-0.75 & +0.35 & -0.65 & +1.05\\ -0.75 & +0.35 & -0.65 & +1.05\\ -0.75 & +0.35 & -0.65 & +1.05\end{pmatrix} + \begin{pmatrix}-0.35 & +0.95 & -0.55 & -0.05\\ +0.40 & -0.40 & -0.10 & +0.10\\ -0.05 & -0.55 & +0.65 & -0.05\end{pmatrix}$$

$$(SS_{Total} = 18.81) = (SS_{Block} = 9.78) + (SS_{Genotype} = 6.63) + (SS_{Residual} = 2.40)$$

For each pair of brackets one may compute the sum of squares (SS) of all entries inside the brackets. The result is shown below the brackets of the above equation. The sum of squares of deviations on the left-hand side of the equation measures the total variation:

$$\begin{aligned}SS_{Total} &= (+0.15)^2 + (+2.55)^2 + (+0.05)^2 + (+2.25)^2 + (-0.75)^2 + (-0.45)^2\\ &\quad + (-1.15)^2 + (+0.75)^2 + (-1.65)^2 + (-1.05)^2 + (-0.85)^2 + (+0.15)^2\\ &= 18.81\end{aligned}$$

The sum of squares for the effects on the right-hand side are computed as follows:

$$\begin{aligned}SS_{Block} &= (+1.25)^2 + (+1.25)^2 + (+1.25)^2 + (+1.25)^2 + (-0.40)^2 + (-0.40)^2\\ &\quad + (-0.40)^2 + (-0.40)^2 + (-0.85)^2 + (-0.85)^2 + (-0.85)^2 + (-0.85)^2\\ &= 9.78\end{aligned}$$

$$\begin{aligned}SS_{Genotype} &= (-0.75)^2 + (+0.35)^2 + (-0.65)^2 + (+1.05)^2 + (-0.75)^2 + (+0.35)^2\\ &\quad + (-0.65)^2 + (+1.05)^2 + (-0.75)^2 + (+0.35)^2 + (-0.65)^2 + (+1.05)^2\\ &= 6.63\end{aligned}$$

$$\begin{aligned}SS_{Residual} &= (-0.35)^2 + (+0.95)^2 + (-0.55)^2 + (-0.05)^2 + (+0.40)^2 + (-0.40)^2\\ &\quad + (-0.10)^2 + (+0.10)^2 + (-0.05)^2 + (-0.55)^2 + (+0.65)^2 + (-0.05)^2\\ &= 2.40\end{aligned}$$

The SS for genotypes measure the difference among genotypes. The larger the differences among genotype means the larger the value of $SS_{Genotype}$. The magnitude of $SS_{Genotype}$ needs to be compared against the residual variation measured by $SS_{Residual}$. This comparison leads to an analysis of variance (ANOVA) table, which has the following format:

Source of variation	Sum of squares SS	Degrees of freedom d.f.	Mean square MS
Blocks	SS_{Block}	$r-1$	
Genotypes	$SS_{Genotype}$	$t-1$	
Error	$SS_{Residual}$	$(r-1)(t-1)$	s^2
Total	SS_{Total}	$rt-1$	

t = no. of genotypes; r = no. of blocks

The mean square is obtained by dividing the sum of squares by its degrees of freedom. The residual mean square is an estimate of the error variance and may be denoted as s^2 for brevity. For the example, we find the following ANOVA table:

Source	SS	d.f.	MS	F
Blocks	9.78	2		
Genotypes	6.63	3	2.21	5.525*
Error	2.40	6	0.40 = s^2	
Total	18.81	11		

The ratio of mean squares for genotypes and error, i.e., the F-value measures the difference among genotypes. The global null hypothesis of no genotype differences can be tested formally by comparing $F = 5.525$ against a tabular F-value at significance level α with 3 numerator d.f. and 6 denominator d.f., which equals $F_{tab} = 5.41$. The F-value is significant at the 5% level of significance.

2.3 Mean comparisons

Following a significant F-test, one will want to compare genotype means. For this purpose, it is necessary to compute the standard error of a difference. For the randomised complete block design this is given by

$$s.e.d. = \sqrt{\frac{2s^2}{r}} \quad ,$$

where s^2 is the residual mean square of the ANOVA and r is the number of blocks. For the example we find

$$s.e.d. = \sqrt{\frac{2 \times 0.40}{3}} = 0.516 \quad .$$

The s.e.d. can be used to compare the difference of two means. The most common method is to compute the least significant difference (LSD) given by

$$LSD = s.e.d. \times t_{tab} \quad ,$$

where t_{tab} is the critical value of a t-distribution with $(t-1)(r-1)$ d.f. at a significance level of 5%. In the case at hand, the tabular t-value for 6 error d.f. is 2.447, so the LSD is found to equal

$$LSD = 0.516 \times 2.447 = 1.263 \text{ t/ha} \; .$$

The LSD can be reported along with the means in a table. In addition, letters can be used to indicate which pairwise comparisons are significant. We will not explain here how the letters are derived. Most computer packages compute such displays.

Genotype	Mean		
1	6.5	a	Means followed by a common letter are not
2	7.6	ab	significantly different at $\alpha = 5\%$ according
3	6.6	a	to the LSD test
4	8.3	b	
LSD	1.26		

In the context of breeding trials, a letter display is not of primary concern, because there are usually many genotypes, making multiple pairwise comparisons rather tedious. Also, the focus is usually on estimation rather than significance testing. Thus, neither the ANOVA F-test nor the LSD-test is itself of primary interest to the breeder. Nevertheless, the LSD is a useful measure of precision and is therefore often reported along with genotype means.

SAS hints

The following code will produce the analysis of variance table and mean comparisons for the example.

```
data e1;
input block cultivar yield;
datalines;
1    1    7.4
1    2    9.8
1    3    7.3
1    4    9.5
2    1    6.5
2    2    6.8
2    3    6.1
2    4    8.0
3    1    5.6
3    2    6.2
3    3    6.4
3    4    7.4
;
proc glm data=e1;
class block cultivar;
model yield=block cultivar;
means cultivar/lsd;
run;

proc mixed data=e1;
class block cultivar;
model yield=block cultivar;
lsmeans cultivar/pdiff;
run;
```

3 Incomplete blocks

So far we have assumed that blocks are complete, i.e., each genotype appears once in each block. Sometimes, data are missing for some plots, so some blocks become incomplete. More importantly, the number of genotypes is often so large, that complete blocks become inefficient as a means of error control due to heterogeneity within blocks. It is then better to use incomplete blocks. There are many designs with incomplete blocks, the most prominent being the flexible class of α**-designs**, which includes so-called **lattice designs** (square lattice and rectangular lattice designs) as special cases. The analysis of incomplete blocks calls for an adjustment of genotype means that accounts for block differences. To motivate this need, we will first look at some artificial examples, where some observations are deleted from complete blocks. The method of least squares is then introduced as a general method for obtaining what are called **adjusted means** or **least squares means** from trials laid out in incomplete blocks. Finally, the method is exemplified for some typical designs with incomplete blocks.

3.1 The need for adjustment

In this section, we will consider data from a randomised complete block design, which are artificially made unbalanced by deleting an observation. The examples will be used to motivate the need for adjustment in computing genotype means.

Example 1: Consider the following hypothetical results of a randomised complete block experiment with three genotypes and three blocks:

		Genotype		
		1	2	3
Block	1	10	20	30
	2	20	30	40
	3	60	70	80
Genotype mean $\bar{y}_{i\bullet}$:		30	40	50

We have simplified the data by assuming that there is no experimental error and that the additive model (1) holds exactly. Thus, differences among genotypes are exactly the same in each block. Genotype 3 has the largest mean and would be judged the best. No one will doubt that for this example the simple mean $\bar{y}_{i\bullet}$ is a useful measure for the performance of a genotype.

Now assume that the observation for genotype 3 in block 3 is missing, i.e., the data are as follows:

		Genotype		
		1	2	3
Block	1	10	20	30
	2	20	30	40
	3	60	70	.
Genotype mean $\bar{y}_{i\bullet}$:		30	40	35

We have computed simple means as before. Due to the missing value, however, the mean for genotype 3 has now changed. Its mean now lies between those for genotypes 1 and 2. Thus, were we to base our assessment on simple means, genotype 3 would be judged differently compared to the balanced case. In fact, the comparison of genotypes would not be fair, because an observation from the most favourable block is missing for genotype 3. Obviously, simple (unadjusted) means are misleading in the case of missing data.

A natural thing to do in order to come up with a fair analysis is to compare genotypes within blocks (**intra-block analysis**). The hypothetical data are such that, e.g., the difference of genotype 3 minus genotype 1 is 20 in blocks 1 and 2. Thus, we would expect the difference to be the same in block 3. We have no observation for genotype 3 in block 3 and thus cannot verify this based on observed data. Instead, we can ask: "What would have been the most likely yield of genotype 3 in block 3?"

Since the yield of genotype 1 is 60, we would expect a yield of 60+20 = 80 for genotype 3. To obtain a corrected or adjusted mean for genotype 3, we can plug in the imputed value into our table and compute simple means, as before:

		Genotype		
		1	2	3
Block	1	10	20	30
	2	20	30	40
	3	60	70	*80*
Genotype mean:		30	40	*50*

imputed value

adjusted mean!

Due to the simplicity of the example, the same result is found for the comparison of genotypes 2 and 3.

Example 2: The first example was rather artificial because absence of experimental error was assumed. Now we add some error by perturbing the data of Example 1 as follows:

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	.
Genotype mean $\bar{y}_{i\bullet}$:		29	41	35

Differences among genotypes now are not the same from block to block due to (simulated) experimental error. An obvious first idea is to analyse the first two complete blocks, discarding the third incomplete block:

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
Genotype mean $\bar{y}_{i\bullet}$:		15	25	35

The mean difference of genotype 3 minus genotype 1 is 20, while the mean difference of genotype 3 minus genotype 2 is 10. The difference between genotype 1 and 2 ist 10. To impute the missing value, this result can be used. The most plausible value for the missing cell in the third block is the one that is in best agreement with the differences computed from the first two blocks. Things are more complicated now, because the difference of genotypes 1 and 2 is 8 in block 3, which is not the same as the difference found from the means based on the first two blocks. An intuitive approach is to plug in a value for the missing cell such that the average of the differences 3–1 and 3–2 is the same in block 3 as that obtained for means from the two first blocks. For the further development it is helpful to denote the imputed value as *m*:

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	*m*

We now compute differences using the variable *m* for the missing value:

	Differences		
	3 minus 1	3 minus 2	average
Means across blocks 1 and 2	20	10	15
Data in block 3	$m - 57$	$m - 73$	$m - 65$

Our requirement was that the average difference in block 3 be the same as the average difference based on means across blocks 1 and 2. Thus, our requirement yields

$15 = m - 65 \Leftrightarrow m = 80$

Using this imputed value to complete the table we find

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	*80*
Genotype mean:		29	41	*50*

imputed value

adjusted mean!

An alternative procedure to impute the missing cell in Example 2 is to consider the four "tetrads" that can be formed between the missing cell and three other cells. A tetrad corresponds to a two-by-two table of means for two cultivars and two blocks. There are four such tables involving block 3 and cultivar 3. One is depicted below.

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	*m*

For clarity, the tetrad is given in the following table:

		Cultivar	
		2	3
Block	2	32	33
	3	73	*m*

The imputed value *m* can be chosen such that the difference of cultivars 2 and 3 is the same in blocks 2 and 3. Thus, we need to solve the equation $(32-33)=(73-m)$, or equivalently the so-called *tetrad-contrast* $(32-33)-(73-m)=0$, yielding $m=-32+33+73=74$ .One may impute the value *m* from each of the four tetrad contrasts. It may be verified that the mean of these different imputed values for *m* equals 80.

We may look at these same computations in a slightly different way, i.e., in a least-squares sense. The four tetrad contrasts are:

$$(12-37)-(57-m)=m-82$$
$$(18-37)-(73-m)=m-92$$
$$(18-33)-(57-m)=m-72$$
$$(32-33)-(73-m)=m-74$$

Ideally, for a tetrad, we would like the tetrad-contrast to equal zero exactly. Obviously we cannot achieve this simultaneously for all tetrads with the same imputed value *m*. What we can do, however, is to try and minimize the departure of a tetrad contrast from zero in a least squares sense. Thus, we would minimize

$$S=(m-82)^2+(m-92)^2+(m-72)^2+(m-74)^2 \ .$$

Equating the first derivative to zero and solving for *m* we find

$$\frac{dS}{dm}=2(m-82)+2(m-92)+2(m-72)+2(m-74)=0 \Leftrightarrow$$
$$m=\frac{82+92+72+74}{4}=80$$

Thus, *m* = 80 is a least squares solution based on tetrads.

We have looked at two simple examples, were we computed "adjusted means" by simple and intuitive algebra. With larger data sets and more than one missing observation, an intuitive approach is difficult to devise and implement. A general and objective method that will yield the same result in the present case, is the method of least squares based on a linear model, to

which we now turn.

3.2 Adjusted means

If we had complete data, we could compute means by taking the simple average

$$\bar{y}_{i\bullet} = \frac{\sum_{i=1}^{r} y_{ij}}{r} \quad .$$

With incomplete data, we may essentially use the same approach after we have imputed the missing data. The imputation may be done by the method of least squares. This method is based on a linear model. For example, the expected value of an observation in the RCB design is

$$\mathrm{E}(y_{ij}) = \eta_{ij} = \mu + b_j + \tau_i \quad .$$

It is natural to define the treatment mean as the simple average of the expected values η_{ij} (rather than observed values y_{ij}) across blocks, i.e.,

$$\bar{\eta}_{i\bullet} = \frac{\sum_{j=1}^{r} \eta_{ij}}{r} \quad .$$

Plugging in the linear model this yields

$$\bar{\eta}_{i\bullet} = \mu + \tau_i + \bar{b}_{\bullet} = \mu + \tau_i + \frac{b_1 + b_2 + \ldots + b_r}{r} \quad .$$

In order to use this formula, we need to compute the least squares estimates of the model parameters, as discussed in the next section. There is no simple scalar formula for these, but they can be easily obtained using a computer. Denoting least squares estimators by the hat-notation as $\hat{\mu}$, $\hat{b}_j$, and $\hat{\tau}_i$, the least squares estimator of the mean $\bar{\eta}_{i\bullet}$ is given by

$$\hat{\bar{\eta}}_{i\bullet} = \hat{\mu} + \hat{\tau}_i + \hat{\bar{b}}_{\bullet} = \hat{\mu} + \hat{\tau}_i + \frac{\hat{b}_1 + \hat{b}_2 + \ldots + \hat{b}_r}{r} \quad .$$

This estimator is also known as the least squares mean or **adjusted mean**. In the case of balanced data, the adjusted mean will coincide with the simple mean (unadjusted mean), i.e. $\hat{\bar{\eta}}_{i\bullet} = \bar{y}_{i\bullet}$. This is why for the randomised complete block design we have computed simple means. For unbalanced data, however, we have $\hat{\bar{\eta}}_{i\bullet} \neq \bar{y}_{i\bullet}$.

3.3 Least squares estimation

Now how do we obtain least squares estimates of the parameters? The method of least squares is well-known from linear regression, where a straight line $\alpha + \beta x$ is fitted to data (y_i, x_i) so

that the sum of squared deviations between observed data y_i and fitted model $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ is minimized:

$$SS_{\text{error}} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left[y_i - \left(\hat{\alpha} + \hat{\beta}x_i\right)\right]^2$$

The least squares estimators are the values of $\hat{\alpha}$ and $\hat{\beta}$ that minimise the residual sum of squares (SS_{Residual}). By analogy, for the randomised complete block design we may minimize

$$SS_{\text{error}} = \sum_{i,j} \left(y_{ij} - \hat{y}_{ij}\right)^2 = \sum_{i,j} \left[y_{ij} - \left(\hat{\mu} + \hat{b}_j + \hat{\tau}_i\right)\right]^2 \quad .$$

We now exemplify least squares estimation for the balanced and unbalanced artificial examples, using an intuitive approach. This is done in order to motivate the method. It is stressed, however, that this intuitive approach is not the one used in actual applications. In practice, a general method of least squares estimation is used, which uses a matrix formulation of linear models. Instead of describing the general method in detail (Searle, 1971), we assume that a computer is available to obtain least squares estimates.

Example 1 (continued):

		Genotype			
		1	2	3	Block mean $\bar{y}_{\bullet j}$
Block	1	10	20	30	20
	2	20	30	40	30
	3	60	70	80	70
Genotype mean $\bar{y}_{i\bullet}$:		30	40	50	40 = overall mean $\bar{y}_{\bullet\bullet}$

For the balanced data, the least squares estimates can be computed as was shown for the analysis of variance in Section 2.1 using the method of sweeping. As is easily verified, the following estimates are identical to those obtained by the method of sweeping:

$\hat{\mu}$ = overall mean = 40
$\hat{\tau}_1$ = mean of genotype 1 – overall mean = 30 – 40 = –10
$\hat{\tau}_2$ = mean of genotype 2 – overall mean = 40 – 40 = 0
$\hat{\tau}_3$ = mean of genotype 3 – overall mean = 50 – 40 = 10
$\hat{b}_1$ = mean of block 1 – overall mean = 20 – 40 = –20
$\hat{b}_2$ = mean of block 2 – overall mean = 30 – 40 = –10
$\hat{b}_3$ = mean of block 3 – overall mean = 70 – 40 = 30

Note that

$\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3 = -10 + 0 + 10 = 0$ and
$\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = -20 - 10 + 30 = 0$

Plugging these estimates into the equation for adjusted means we find

$$\hat{\bar{\eta}}_{1\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_1 = 40 + \frac{-20-10+30}{3} - 10 = 30$$
$$\hat{\bar{\eta}}_{2\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_2 = 40 + \frac{-20-10+30}{3} + 0 = 40$$
$$\hat{\bar{\eta}}_{3\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_3 = 40 + \frac{-20-10+30}{3} + 10 = 50$$

These least squares means are identical to simple means because the data were balanced. The computations we have just shown only work for balanced data. More generally, the least squares solutions of the parameters are obtained using a different method. Using a computer which employs this more general method, we find the following least squares solution for the parameters for the same balanced data set:

$$\begin{pmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{pmatrix} = \begin{pmatrix} 80 \\ -50 \\ -40 \\ 0 \\ -20 \\ -10 \\ 0 \end{pmatrix}$$

Note that the last genotype effect and the third block effect are equal to zero. These solutions do not obey the sum-to-zero restrictions $\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3 = 0$ and $\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = 0$. Plugging these estimates into the equations for least squares means we find

$$\hat{\bar{\eta}}_{1\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_1 = 80 + \frac{-50-40+0}{3} - 20 = 30$$
$$\hat{\bar{\eta}}_{2\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_2 = 80 + \frac{-50-40+0}{3} - 10 = 40$$
$$\hat{\bar{\eta}}_{3\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_3 = 80 + \frac{-50-40+0}{3} + 0 = 50$$

These are exactly the same means as before, even though the least squares solutions of effects were different! Obviously, there is more than one way of computing least squares solutions. The solutions are said to be non-unique. Because of the non-uniqueness, some kind of restriction needs to be imposed on the parameters. In the first case the restrictions were $\hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3 = 0$ and $\hat{b}_1 + \hat{b}_2 + \hat{b}_3 = 0$, and these are often used in textbooks. In the second case the restrictions are $\hat{\tau}_3 = 0$ and $\hat{b}_3 = 0$, which are used by most computer packages. With either restriction, we obtain the same adjusted means.

Example 2 (continued): We now consider the second unbalanced dataset.

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	.

The following least squares solutions are found using a computer:

$$\begin{pmatrix} \hat{\mu} \\ \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{pmatrix} = \begin{pmatrix} 80 \\ -47.66 \\ -42.33 \\ 0 \\ -21 \\ -9 \\ 0 \end{pmatrix}$$

$$\hat{\bar{\eta}}_{1\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_1 = 80 + \frac{-47.66 - 42.33 + 0}{3} - 21 = 29$$

$$\hat{\bar{\eta}}_{2\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_2 = 80 + \frac{-47.66 - 42.33 + 0}{3} - 9 = 41$$

$$\hat{\bar{\eta}}_{3\bullet} = \hat{\mu} + \frac{\hat{b}_1 + \hat{b}_2 + \hat{b}_3}{3} + \hat{\tau}_3 = 80 + \frac{-47.66 - 42.33 + 0}{3} + 0 = 50$$

These are the same means as obtained with the intuitive method, by which we first imputed the missing value and then computed simple means for the completed two-way table.

SAS hints

```
data toy;
input block genotype yield;
datalines;
1 1 12
1 2 18
1 3 37
2 1 18
2 2 32
2 3 33
3 1 57
3 2 73
;
proc glm data=toy;
class block genotype;
model yield=block genotype / solution; /*-least squares estimates of
effects*/
means genotype; /*arithmetic means*/
lsmeans genotype; /*adjusted means*/
run;
```

3.4 Recovery of inter-block information and mixed modelling

An important aspect of the least squares estimation described in Sections 3.2 and 3.3 is that all information on genotypes is obtained from comparisons among genotypes within blocks. This information is denoted as **intra-block information**. The least squares method described in Sections 3.2 and 3.3 is therefore also known as **intra-block analysis**.

Example 3: To further elucidate intra-block analysis, we use blocks 2 and 3 of Example 2 and delete one further observation.

		Genotype		
		1	2	3
Block	2	18	.	33
	3	57	73	.

There are three comparisons of interest: 1 vs. 2, 1 vs. 3, and 2 vs. 3. Each of the last two comparisons can be based on information of one block:

Difference "genotype 1 – genotype 2" = 57 – 73 = –16 (observed in block 3)
Difference "genotype 1 – genotype 3" = 18 – 33 = –15 (observed in block 2)

Both of these differences obviously exploit **direct comparisons** from within a single block. Now what about the comparison of genotypes 2 and 3? These are not observed in the same block, so a direct comparison is not possible. It is easily seen, however, that this difference can be indirectly inferred from the two direct differences, as

Difference "genotype 2 – genotype 3" = Difference "genotype 1 – genotype 3" in block 2
– Difference "genotype 1 – genotype 2" in block 3
= –15 –(–16)
= +1

This difference is called an **indirect comparison**.

The direct and indirect comparisons may now be compared to the least squares estimates. Using a computer, we find the following least squares solutions for the model effects:

$$\begin{pmatrix} \hat{\mu} \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{\tau}_1 \\ \hat{\tau}_2 \\ \hat{\tau}_3 \end{pmatrix} = \begin{pmatrix} 72 \\ -39 \\ 0 \\ -15 \\ 1 \\ 0 \end{pmatrix}$$

The differences among genotype effects are identical to the two direct differences and the one indirect comparison computed above. They are also identical to the corresponding differences among adjusted means:

$$\hat{\bar{\eta}}_{1\bullet} = \hat{\mu} + \frac{\hat{b}_2 + \hat{b}_3}{2} + \hat{\tau}_1 = 72 + \frac{-39+0}{2} - 15 = 37.5$$
$$\hat{\bar{\eta}}_{2\bullet} = \hat{\mu} + \frac{\hat{b}_2 + \hat{b}_3}{2} + \hat{\tau}_2 = 72 + \frac{-39+0}{2} + 1 = 53.5$$
$$\hat{\bar{\eta}}_{3\bullet} = \hat{\mu} + \frac{\hat{b}_2 + \hat{b}_3}{2} + \hat{\tau}_3 = 72 + \frac{-39+0}{2} + 0 = 52.5$$

For example,

$\hat{\tau}_1 - \hat{\tau}_2 = -15 - 1 = -16$
$\hat{\eta}_1 - \hat{\eta}_2 = 37.5 - 53.5 = -16$
Difference "genotype 1 – genotype 2" = 57 – 73 = –16 (observed in block 3)

This example shows that least squares estimation exploits the intra-block information derived from direct and indirect differences. In addition, some information can also be obtained from block sums, as will be shown now. The information thus exploited is denoted as **inter-block information**.

Example 3 (continued):

We may compute

Sum "genotype 1 + genotype 3" in block 2 = 51 and
Sum "genotype 1 + genotype 2" in block 3 = 130.

Obviously, the difference of the two sums provides an estimate of the difference of genotypes 2 and 3:

Difference "genotype 2 – genotype 3" = Sum "genotype 1 + genotype 2" in block 3
- Sum "genotype 1 + genotype 3" in block 2
= 130 - 51
= 79

This difference exploits what is known as **inter-block information**. In this case its value happens to be rather different from the indirect difference, which equalled +1. This latter difference is based on the intra-block information.

Now that we have two different estimates of the difference, how should we combine them into a single value? One could compute a simple average of the two, but this is not usually the best strategy. It is often the case that the inter-block information is much less accurate than the intra-block information. Thus, the intra-block information should be given more weight. Accuracy is best assessed in terms of variance: the smaller the variance the higher the accuracy. Therefore, the inverse of the variance of a difference can be used as a weight in a weighted mean of the two difference estimates, as we will see shortly.

In order to assess the variance of the two difference estimates, it is necessary to reconsider the linear model for blocked experiments. For incomplete blocks, we may use the same model as

with complete blocks, i.e.,

$$y_{ij} = \mu + b_j + \tau_i + e_{ij} \quad ,$$

where y_{ij} is the observed yield of the i-th genotype in the j-th block, b_j is the effect of the j-th block, τ_i is the effect of the i-th genotype, and e_{ij} is the residual effect corresponding to y_{ij}. It is now necessary to make some more specific assumptions regarding the block and error effects. Both of these effects will be assumed to be random effects with variances

$$\text{var}(e_{ij}) = \sigma_e^2 \quad \text{and}$$
$$\text{var}(b_j) = \sigma_b^2 \quad .$$

With these assumptions our model becomes a **mixed model**. A mixed model is a linear model, in which more than one effect is random, and in addition, there is at least one fixed effect apart from the general mean.

We can now determine the variances of the two difference estimates. This is best done by using the model in the expression for the estimator.

Example 3 (continued): We find for the indirect comparison (intra-block information)

D_1 = Difference "genotype 2 – genotype 3" = Difference "genotype 1 – genotype 3" in block 2
– Difference "genotype 1 – genotype 2" in block 3

$$\begin{aligned}
&= y_{12} - y_{32} \\
&- (y_{13} - y_{23}) \\
&= \mu + b_2 + \tau_1 + e_{12} - (\mu + b_2 + \tau_3 + e_{32}) \\
&- [(\mu + b_3 + \tau_1 + e_{13}) - (\mu + b_3 + \tau_2 + e_{23})] \\
&= \tau_2 - \tau_3 + (e_{12} - e_{32} - e_{13} + e_{23})
\end{aligned}$$

The first thing to notice is that a number of effects cancel out, and this should not be surprising, as we are looking at a difference of a difference. Specifically, the block effect drops out, essentially because we exploit intra-block information by differencing within blocks. Thus, the variance of D_1 only depends on the variance of e_{ij}. There are four error effects involved, so that

$$\text{var}(D_1) = 4\sigma_e^2 \quad .$$

For the inter-block information, we find

D_2 = Difference "genotype 2 – genotype 3" = Sum "genotype 1 + genotype 2" in block 3
– Sum "genotype 1 + genotype 3" in block 2

$$\begin{aligned}
&= y_{13} + y_{23} \\
&- (y_{12} + y_{32}) \\
&= [(\mu + b_3 + \tau_1 + e_{13}) + (\mu + b_3 + \tau_2 + e_{23})] \\
&- [(\mu + b_2 + \tau_1 + e_{12}) + (\mu + b_2 + \tau_3 + e_{32})] \\
&= \tau_2 - \tau_3 + 2b_3 - 2b_2 + (e_{13} + e_{23} - e_{12} - e_{32})
\end{aligned}$$

The variance of this difference depends on both the block effects and the error effects.

Specifically, there are two block effects, and they enter with a coefficient of two. As for the intra-block information, there are four error effects. Thus, the variance is

$$\mathrm{var}(D_2) = 2 * 2^2 \sigma_b^2 + 4\sigma_e^2 = 8\sigma_b^2 + 4\sigma_e^2 \quad .$$

This result shows that the inter-block difference D_2 is generally less accurate than the intra-block difference D_1. One may combine the two differences by a weighted mean, with weights of a difference equal to the inverse of the variance:

$$D = \frac{w_1 D_1 + w_2 D_2}{w_1 + w_2} \quad ,$$

where

$$w_1 = \frac{1}{\mathrm{var}(D_1)} \quad \text{and}$$
$$w_2 = \frac{1}{\mathrm{var}(D_2)} \quad .$$

This estimator is known as the combined **inter-block-intra-block estimator**. It can be shown to have the smallest variance among all weighted linear estimators. This property is also known as **BLUE**: best linear unbiased estimator. In order use this estimator in practice, we need to estimate the variances, which determine the weights. Here, we will not discuss estimation, but assume that the variances are

$$\sigma_e^2 = 1$$
$$\sigma_b^2 = 5$$

With these values we find

$\mathrm{var}(D_1) = 4$, $w_1 = 1/4 = 0.25$ $\quad [D_1 = 1]$
$\mathrm{var}(D_2) = 44$, $w_2 = 1/44 = 0.022727$ $\quad [D_2 = 79]$

and for the combined **inter-block-intra-block estimator**

$$D = \frac{1 \times 0.25 + 79 \times 0.022727}{0,.25 + 0.022727} = 7.5 \quad .$$

It is seen that this estimate is dominated by the intra-block information, which is a result is the much higher weight for this information compared to the inter-block information.

SAS hints

To obtain the combined intra-bloc-inter-block analysis, all we need to do is to fit the block effect as random. This is done by declasing block as a random effect using the RANDOM statement.

To reproduce the computations for the toy example, where we had to fix the variance components, you need to use the PARMS statement as shown below. The HOLD= option is used to fix the parameters at their starting values and this prevent iterations of the algorithm that would otherwise estimate the variance components from the data. – With a larger dataset that allows estimating the variances from the data, simply drop the PARMS statement.

```
data a;
input block geno yield;
datalines;
2 1 18
2 3 33
3 1 57
3 2 73
;
proc mixed data=a;
class block geno;
model yield=geno/solution;
random block;
parms (5)(1)/hold=1,2;
lsmeans geno/pdiff;
run;
```

***Why is *D* really best?**

$$D = \frac{w_1 D_1 + w_2 D_2}{w_1 + w_2} = v_1 D_1 + v_2 D_2 ,$$

where

$$v_1 = \frac{w_1}{w_1 + w_2} \text{ and } v_2 = \frac{w_2}{w_1 + w_2}$$

We have $v_1 + v_2 = 1 \Leftrightarrow v_2 = 1 - v_1$. Thus

$D = v_1 D_1 + (1 - v_1) D_2$ and therefore

$\text{var}(D) = v_1^2 \text{ var}(D_1) + (1 - v_1)^2 \text{ var}(D_2)$.

We seek to minimize this variance, i.e., we seek the value of v_1 that minimizes $\text{var}(D)$. This optimum is found by computing the first derivative of $\text{var}(D)$ with respect to v_1 and then solving for v_1.

$$\frac{\partial \text{var}(D)}{\partial v_1} = 2v_1 \text{ var}(D_1) + 2(1 - v_1)(-1)\text{var}(D_2) = 0 \Leftrightarrow$$
$$v_1[\text{var}(D_1) + \text{var}(D_2)] = \text{var}(D_2) \Leftrightarrow$$
$$v_1 = \frac{\text{var}(D_2)}{\text{var}(D_1) + \text{var}(D_2)}$$

Now show that $v_1 = \dfrac{w_1}{w_1 + w_2}$: Divide both numerator and denominator by $\text{var}(D_1) \times \text{var}(D_2)$:

$$v_1 = \frac{\text{var}(D_2)}{\text{var}(D_1) + \text{var}(D_2)} = \frac{\dfrac{\text{var}(D_2)}{\text{var}(D_1) \times \text{var}(D_2)}}{\dfrac{\text{var}(D_1) + \text{var}(D_2)}{\text{var}(D_1) \times \text{var}(D_2)}} = \frac{\dfrac{1}{\text{var}(D_1)}}{\dfrac{1}{\text{var}(D_1)} + \dfrac{1}{\text{var}(D_2)}} = \frac{w_1}{w_1 + w_2} \quad \text{(q.e.d.)}$$

***3.5 Estimation of variance components by ML and REML: a simple example**

The standard method for estimating variance components is the restricted maximum likelihood (REML) method, a variant of the maximum likelihood (ML) method that is often preferred for variance component estimation. The method seeks to find values of the variance components that give the greatest probability or likelihood to the observed data under the assumed model. Maximizing the likelihood usually needs to be done by iterative methods, except in very simple cases.

This is not the place to give a detailed account of the REML method, which may be found elsewhere (Searle et al., 1992). Here, we just give a brief characterization and motivation based on the example of a simple random sample. We first exemplify the idea of maximum likelihood (ML) estimation for the mean and contrast this to the method of least squares. We then consider estimation of the variance by REML and ML.

Assume that a variance is to be estimated for a simple random sample of values $y_1, y_2, \ldots, y_n$ assumed to obey the model

$$y_i = \mu + e_i \quad ,$$

where

μ = expected value
e_i = error of i-th value

Estimation of the mean by least squares: The method of least squares seeks to find the value of the mean parameter μ that minimizes the sum of squared deviations between data and fitted mean. This, the objective function to be minimized is the error sum of squares (SS_{error}) given by

$$SS_{error} = \sum_{i=1}^{n} (y_i - \mu)^2 \quad .$$

To find the minimum with respect to μ, we compute the first derivative with respect to μ, equate to zero and solve for μ. The first derivative is

$$\frac{\partial SS_{error}}{\partial \mu} = \sum_{i=1}^{n} -2(y_i - \mu) \quad .$$

Equating this to zero, we find

$$\sum_{i=1}^{n} -2\left(y_j - \mu\right) = 0 \Leftrightarrow n\mu = \sum_{i=1}^{n} y_i \quad .$$

So the least squares solution is

$$\hat{\mu}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_{\bullet} \ .$$

This solution really minimizes the error sum of squares, because the second derivative of SS_{error} at $\hat{\mu}_{LS}$ is positive:

$$\frac{\partial^2 SS_{error}}{\partial \mu^2} = 2n > 0 \quad .$$

Summary: The method of least squares is one of two very general estimation principles in statistics. The simple sample mean is the least squares estimator of the mean, giving a nice justification for this very common estimator.

Estimation of mean and variance by Maximum Likelihood (ML): ML estimation is based on the **probability density** of the data. Assuming a normal distribution, for a single observation this is given by

$$f\left(y_i, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \ .$$

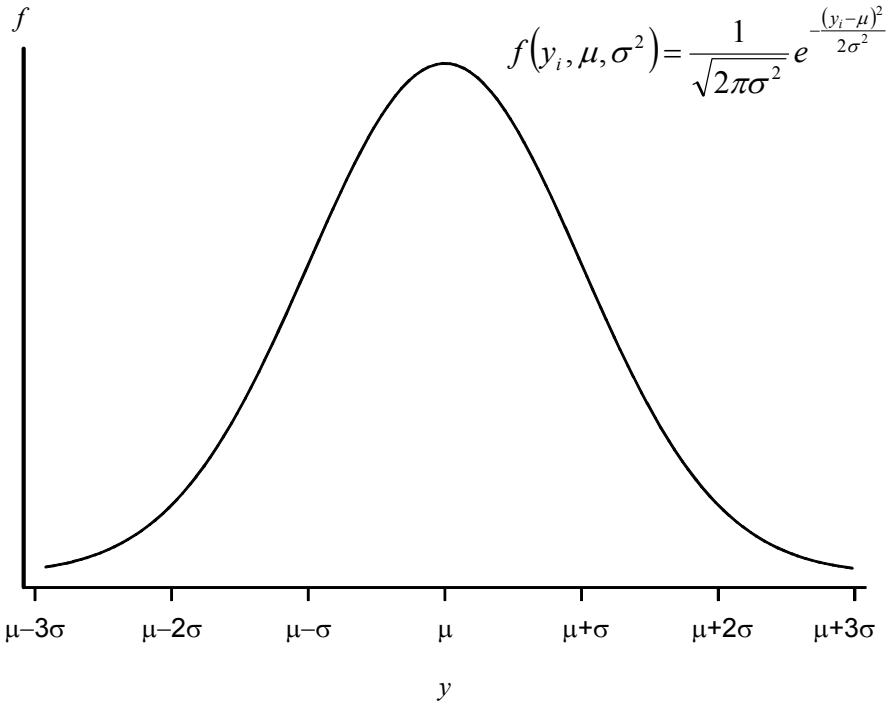


Fig. 2: Probability density of normal distribution.

The largest density is at $y_i = \mu$, and the density drops symmetrically on both sides (Fig. 2). The larger the variance σ^2, the slower is the drop in density. The joint density for a sample of independent observations is obtained by multiplying the densities for the individual observations. One may multiply densities because observations y_i are independent. The joint density is

$$f(\boldsymbol{y}, \mu, \sigma^2) = f(y_1, \mu, \sigma^2) \times f(y_2, \mu, \sigma^2) \times .... \times f(y_n, \mu, \sigma^2),$$

giving the probability density of y_j for given values of μ and σ^2 (Fig. 2). This interpretation considers parameters μ and σ^2 as fixed and known, while observations y_i are variable.

The density may also be used to estimate the parameters. This is done by changing the perspective: The data y_i are taken as fixed quantities, while parameters become variables. We may then ask: which values of the parameters give the largest probability density to the observed data y_i? In other words, which values of the parameters maximize the likelihood of y_i? The parameter values maximizing the probability density, or likelihood, are in a certain sense the most plausible parameter values, given the observed data. When this perspective is taken, the probability density is called the **likelihood**. And the parameter values maximizing the likelihood for given data y_j are called the **Maximum Likelihood** (ML) estimates or estimators.

For ML estimation we must consider the likelihood of the complete dataset, i.e.

$$L = f(\boldsymbol{y}, \mu, \sigma^2) = f(y_1, \mu, \sigma^2) \times f(y_2, \mu, \sigma^2) \times .... \times f(y_n, \mu, \sigma^2) \ ,$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ is the vector of observations y_i. Maximizing the likelihood is easier when taking the logarithm of the likelihood, the so-called **log-likelihood**:

$$\log L = \log\left[f(\boldsymbol{y}, \mu, \sigma^2)\right] = \log\left[f(y_1, \mu, \sigma^2)\right] + \log\left[f(y_2, \mu, \sigma^2)\right] + ... \log\left[f(y_n, \mu, \sigma^2)\right]$$

The values of the parameters maximizing the likelihood L will also maximize the log-Likelihood (log L). Assuming normality we have

$$\log\left[f(y_i, \mu, \sigma^2)\right] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mu)^2}{2\sigma^2}$$

and thus

$$\log L = \sum_{i=1}^{n}\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mu)^2}{2\sigma^2}\right\}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\left\{\frac{(y_i - \mu)^2}{2\sigma^2}\right\}.$$

In order to find the maximum, we compute partial derivatives of log L with respect to μ and σ^2, equate these to zero and solve for the parameters:

$$\frac{\partial \log L}{\partial \mu} = -2\sum_{i=1}^{n}\frac{(y_i-\mu)}{\sigma^2} = 0 \Leftrightarrow \sum_{i=1}^{n} y_i = n\mu \Leftrightarrow \hat{\mu}_{ML} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}_\bullet$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2(\sigma^2)^2} = 0 \Leftrightarrow n\sigma^2 = \sum_{i=1}^{n}(y_i-\mu)^2 \Leftrightarrow \sigma^2 = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n} \quad .$$

It turns out that the sample mean is also an ML estimator of the mean μ. In other words, the sample mean is that value for the mean μ, which gives the observed data the largest likelihood or, in less statistical terms, the largest plausibility. Plugging in the ML estimator for μ in the estimating equation for σ^2 yields

$$\hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^{n}(y_i-\bar{y}_\bullet)^2}{n} \quad .$$

An important point to note here is that the error sum of squares is not divided by $n-1$ but by n. For this reason, ML estimation does not yield an unbiased estimator. We have

$$E(\hat{\sigma}^2_{ML}) = \frac{n-1}{n}\sigma^2 < \sigma^2 \quad .$$

The problem here is that the degrees of freedom are not accounted for. To obtain an unbiased estimator we need to divide by $n-1$, because one degree of freedom is lost by estimating the mean. Thus, the usual estimator is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y}_\bullet)^2}{n-1} \quad .$$

We will come back to this issue when we discuss the REML method for this example. The maximized log-likelihood is

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\hat{\sigma}^2_{ML}\right) - \sum_{j=1}^{n}\left\{\frac{(y_j - \hat{\mu})^2}{2\hat{\sigma}^2_{ML}}\right\}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sum_{j=1}^{n}(y_j - \bar{y}_\bullet)^2}{n}\right) - \frac{1}{2\frac{\sum_{j=1}^{n}(y_j - \bar{y}_\bullet)^2}{n}}\sum_{j=1}^{n}(y_j - \bar{y}_\bullet)^2 .$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\sum_{j=1}^{n}(y_j - \bar{y}_\bullet)^2}{n}\right) - \frac{n}{2}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{SS_{error}}{n}\right) - \frac{n}{2}$$

Interestingly, for this simple example this depends on the data only through the error sum of squares.

***Summary*:** The method of Maximum likelihood is one of two very general estimation principles in statistics. The usual sample mean can be justified as maximum likelihood estimator of the population mean μ. The ML estimator of the variance σ^2 is biased. The maximized likelihood for a simple random sample depends on the data only through the residual error sum of squares.

REML method: The Restricted Maximum Likelihood (REML) method also is an ML method. It accounts for the loss of degrees of freedom from estimating fixed effects. This is achieved by using contrasts that are free of fixed effects. In the present case there is just a single effect μ. A contrast is a linear combination of the data such that the expected value is zero. One such contrast is

$$z_1 = \frac{1}{\sqrt{2}}y_1 - \frac{1}{\sqrt{2}}y_2 ,$$

because

$$E(z_1) = 0 \text{ and } \text{var}(z_1) = \sigma^2 .$$

A further contrast is

$$z_2 = \frac{1}{\sqrt{6}}y_1 + \frac{1}{\sqrt{6}}y_2 - \frac{2}{\sqrt{6}}y_3$$

with

$$E(z_2) = 0 \text{ and } \text{var}(z_2) = \sigma^2 .$$

These two contrasts are orthogonal and therefore independent. We can form n-1 such contrasts. The coefficients can be compiled in a matrix $\boldsymbol{H}$ with $n-1$ rows and n columns:

$$\boldsymbol{H} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & & & \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & & \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & -\frac{3}{\sqrt{12}} & \\ \vdots & \vdots & \vdots & \vdots & \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & . & . & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

The contrasts are given by

$$\boldsymbol{z} = (z_1, z_2, ..., z_{n-1})' = \boldsymbol{Hy} \quad ,$$

with $E(\boldsymbol{z}) = \boldsymbol{0}$ and $\mathrm{var}(\boldsymbol{z}) = \sigma^2 \boldsymbol{I}_{n-1}$. The contrasts are orthogonal, because $\boldsymbol{HH'} = \boldsymbol{I}_{n-1}$. Orthogonality means that the contrasts z_j are independent. Bacause of the independence, we can write the likelihood for $\boldsymbol{z}$ as a product of the likelihoods of z_j. Importantly, the contrasts are free of the fixed effect μ, so they contain information only pertaining to the variance, but not the mean.

The likelihood of contrasts z_j, the so-called restricted log-likelihood, is

$$\log L = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log(\sigma^2) - \sum_{j=1}^{n-1}\left\{\frac{z_j^2}{2\sigma^2}\right\} \quad .$$

The restricted log-likelihood only depends on the variance, but not on the mean, which has been removed by the contrasts. We find

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n-1}{2\sigma^2} + \sum_{j=1}^{n-1}\frac{z_j^2}{2(\sigma^2)^2} = 0 \Leftrightarrow \hat{\sigma}^2_{REML} = \frac{\sum_{j=1}^{n-1} z_j^2}{n-1} \quad .$$

It can be shown that

$$\sum_{j=1}^{n-1} z_j^2 = \boldsymbol{z'z} = \boldsymbol{y'H'Hy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = \sum_{i=1}^{n}(y_i - \bar{y}_\bullet)^2$$

so that

$$\hat{\sigma}^2_{REML} = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_\bullet)^2}{n-1} = s^2 \, .$$

This shows that the restricted ML (REML) estimator is equal to the usual variance estimator for a simple random sample, which is known unbiased.

In general, REML as applied to more complex models, including one for incomplete blocks, does not entirely remove bias, but usually reduces bias compared to ML. For this reason, REML is preferred to ML (Searle et al., 1992). The method can be applied to any mixed linear model, for example the model for an incomplete block design. In general, the REML method uses $n-p$ linearly independent contrasts, when the model has p fixed effects and n observations. In case of a linear model for t fixed genotypes and r random blocks, there are t fixed effects to be removed, so $n-t$ contrasts can be formed. The resulting restricted log-likelihood is somewhat more elaborate and will not be given here. Suffice it to say that REML is the standard method for estimating variance components. A more general treatment of REML will be given in Section 4.2.

3.6 Intra-block and combined inter-intra-block analysis using a mixed model package

It should be stressed that the assumption of random block effects was needed to obtain a variance of the inter-block difference D_2. By contrast, the intra-block analysis does not require the assumption of random blocks. In fact, in order to perform an intra-block analysis, one will use a linear model package taking block effects as fixed. Conversely, if a mixed model package is used to fit a model with random blocks, the resulting mean estimates for the genotypes are equivalent to the combined inter-block-intra-block analysis described above.

Exploiting the inter-block information is only worthwhile if the block variance can be estimated with good accuracy. A rule of thumb says that the number of blocks needs to be at least 10.

Having discussed the basic principles of the analysis of incomplete block designs, using toy datasets, we will now consider several real examples, assuming that a mixed model package is available.

3.7 Analysis of a non-resolvable augmented design

An augmented design in its simplest form consists of a common design for several standard genotypes, such as the randomised complete block design. The blocks are augmented in order to accommodate new entries (genotypes, lines, etc.). Each entry is tested in only one block. In effect, the standards are replicated while the new entries are not. The design is very useful when there are very many entries and the trial capacity is limited. It is also useful in early-generation testing, when the plant material per entry is limited.

Example 4: An augmented design with three standards (st, ci, wa) and 30 entries was laid out with 6 blocks. (Pattersen, 1994) (**augmented.dat**).

Block I		Block II		Block III	
Genotype	Yield	Genotype	Yield	Genotype	Yield
st	2972	st	3122	st	2260
14	2405	ci	3023	18	2603
26	2855	4	3018	27	2857
ci	2592	15	2477	ci	2918
17	2572	30	2955	25	2825
wa	2608	3	3055	28	1903
22	2705	wa	2477	5	2065
13	2391	24	2783	wa	3107

Block IV		Block V		Block VI	
Genotype	Yield	Genotype	Yield	Genotype	Yield
st	3348	st	1315	st	3538
9	2268	2	1055	29	2915
6	2148	21	1688	7	3265
ci	2940	wa	1625	ci	3483
wa	2850	ci	1398	1	3013
20	2670	10	1293	wa	3400
11	3380	8	1253	12	2385
23	2770	16	1495	19	3643

The three standards are tested in each block, while each entry is tested in only one of the blocks. This design gives rise to incomplete blocks, so the augmented design is just one special form of an incomplete block design. Thus, we analyse the data based on the usual model $y_{ij} = \mu + b_j + \tau_i + e_{ij}$. There are only six blocks, so the number of observations for estimating the block variance is too small for an efficient recovery of inter-block information (see rule of thumb in Section 3.6). Consequently, we only perform an intra-block analysis. The least squares means are:

Entry	adj. mean
1	2260.22
2	2329.89
3	2901.89
4	2864.89
5	2024.22
6	1822.89
7	2512.22
8	2527.89
9	1942.89
10	2567.89
11	3054.89
12	1632.22

13	2387.89
14	2401.89
15	2323.89
16	2769.89
17	2568.89
18	2562.22
19	2890.22
20	2344.89
21	2962.89
22	2701.89
23	2444.89
24	2629.89
25	2784.22
26	2851.89
27	2816.22
28	1862.22
29	2162.22
30	2801.89

Standard	adj. mean
ci	2725.67
st	2759.17
wa	2677.83

The standard error of a difference (s.e.d.) among two entries depends on whether or not the two entries were tested in the same block or in different blocks. In the first case, we have a direct difference, while in the second case we have an indirect comparison via the standards, which is less accurate. If we denote the standards as S1, S2, and S3 and two entries in the same block as E1 and E2, then a direct difference can be expressed as

$D = E1\text{-}E2$,

which has variance $2\sigma^2$. In case of an indirect difference (entries E1 and E2 are in different blocks), we can compute the difference to the mean of the standards of a block and that compute the difference of these differences for E1 and E2:

$$E1 - \frac{S1+S2+S3}{3} \text{ in block 1} \quad \text{minus} \quad E2 - \frac{S1+S2+S3}{3} \text{ in block 2,}$$

which has variance $\frac{8}{3}\sigma^2$. This is larger than the variance of a direct difference. The s.e.d.'s, which are the square-root s of these variances, are s.e.d. = 427 for the direct difference and 493 for the indirect difference. Both s.e.d.'s are quite large relative to the range of means, so the numerical differences among means should not be over-interpreted.

As we have two different s.e.d., there is no common s.e.d. or LSD to report here. This situation is typical for analysis of designs with incomplete blocks.

Recovery of inter-block information is not worthwhile. This becomes apparent, if we use the Kenward-Roger method to approximate the s.e.d. under a model with random block effects.

This method accounts for the fact that variance components are not known but need to be estimated, which adds variability to the estimates of linear effects.

Analysis	s.e.d. for entries
Block fixed	483.78
Blocks random	480.30
Blocks random with Kenward-Roger	484.48

The s.e.d. is slightly smaller for a model with fixed block effects than for the model with random effects, when estimation error is accounted for (Kenward-Roger method), confirming that recovery of inter-block information is not worthwhile.

SAS hints

In order to implement the Kenward-Roger method, which involves both a correction of the standard errors as well as an adjustment of the denominator degrees of freedom, you can use the option DDFM=KR to the MODEL statement. The mean s.e.d. can be used by outputting the differenced and associated standard errors using the ODS statement (ODS=output delivery system). Standard errors can then be averaged using the MEANS procedure. In the above table, I averaged s.e.d. only for comparisons that do not involve the checks. In most other applications, one will just average all s.e.d.s Thus, I am here giving just the code to compute that overall average (which will produce other values than the ablve table, but the same ranking of methods). Also, I am only giving the code for (i) fixed blocks and (ii) random blocks in combination with the Kenward-Roger method, because these are the two methods we should be comparing in practice.

```
data aug;
input block cultivar$ yield;
datalines;
1     st    2972
1     14    2405
1     26    2855
<more data>
6     wa    3400
6     12    2385
6     19    3643
;
/*blocks fixed*/
ods output diffs=sed_fixed_blocks;
proc mixed data=aug;
class block cultivar;
model yield=cultivar block;
lsmeans cultivar/pdiff;
run;

proc means data=sed_fixed_blocks mean;
var StdErr;
run;

/*random blocks with Kenward-Roger method*/

ods output diffs=sed_random_blocks_with_KR;
proc mixed data=aug;
class block cultivar;
model yield=cultivar / ddfm=KR; /*Kenward Roger method*/
random block;
lsmeans cultivar/pdiff;
run;
```

```
proc means data=sed_random_blocks_with_KR mean;
var StdErr;
run;
```

3.8 Analysis of an α-design

It is often desirable that incomplete blocks can be grouped to form complete replicates, meaning that each treatment occurs exactly once in a replicate. Such designs are called **resolvable**. Resolvability allows differences between replicates to be separated from error and also from the block variance. For example, it may not be possible to harvest a large field trial on a single day, but it may be feasible to havest a complete replicate in a single day. In this case, a resolvable design is very useful, if harvest is done one replicate a day. In this case, larger differences may be expected between replicates, and these can be accounted for by fitting a replicate effect. Without such a replicate effect, the variance of incomplete block effects would be inflated due to differences between harvest days, adversely affecting the efficiency of the design.

The most flexible class of resolvable incomplete block designs are **α-designs** proposed by Emlyn Williams and Desmond Patterson in the 1970s, which can be generated from so-called α-arrays (John & Williams, 1995). For block size to be constant for an α-design, the number of treatments must be a multiple of the block size, i.e. the number of treatments must be

$v = sk$,

where

s = number of blocks per replicate
k = block size
v = number of treatments.

There are some special cases of re-solvable incomplete block designs that were proposed by Frank Yates in the 1940s. So-called **square lattice designs** require the number of treatments to be an exact square $v = k^2$, and block size to be the square root of the number of treatments, such that $s = k$. **Rectangular lattice designs** require $v = k(k+1)$ and $s = k+1$. These designs can usually be obtained as special cases of an α-design and so are included in the broader class of α-designs. Lattice designs were of great importance when they were first introduced, and they continue to play an important role. A practical limitation of lattice designs is the restriction on treatment number and block size. For this reason, we focus our attention on α-designs, which provide much more flexibility, though occasionally we make specific reference to lattice designs in examples and exercises.

Example 5: John and Williams (1995) report results from a yield trial with oats laid out as an α-design. The trial had 24 genotypes, three complete replicates and six incomplete blocks within each replicate. The block size was four.

The design was as follows:

Replicate 1	Replicate 2	Replicate 3
		Block no.

1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
11	21	23	13	17	6	8	24	12	5	2	19	11	2	17	12	21	3
4	10	14	3	15	12	20	15	11	9	18	7	1	15	18	13	22	5
5	20	16	19	7	24	14	3	21	10	13	6	14	9	4	10	16	20
22	2	18	8	1	9	4	23	17	1	22	16	19	8	6	23	24	7

The yields in t/ha were as follows (**alpha.dat**):

Replicate 1

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
4.1172	4.6540	4.2323	4.2530	4.7876	4.7085
4.4461	4.1736	4.7572	3.3420	5.0902	5.2560
5.8757	4.0141	4.4906	4.7269	4.1505	4.9577
4.5784	4.3350	3.9737	4.9989	5.1202	3.3986

Replicate 2

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
3.9926	3.9039	5.3127	5.1202	5.1566	5.3148
3.6056	4.9114	5.1163	4.2955	5.0988	4.6297
4.5294	3.7999	5.3802	4.9057	5.4840	5.1751
4.3599	4.3042	5.0744	5.7161	5.0969	5.3024

Replicate 3

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
3.9205	4.0510	4.3234	4.1746	4.4130	2.8873
4.6512	4.6783	4.2486	4.7512	4.2397	4.1972
4.3887	3.1407	4.3960	4.0875	4.3852	3.7349
4.5552	3.9821	4.2474	3.8721	3.5655	3.6096

An α-design is a design with incomplete blocks, were the blocks can be grouped into complete replicates. Such designs are termed "**resolvable**". The model must have an effect for complete replicates, and effects for incomplete blocks must be nested within replicates. The model is

$$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + e_{ijh}$$

where

y_{ijh} = yield of i-th genotype in h-th block nested within j-th complete replicate
μ = general effect
γ_j = effect of j-th complete replicate
b_{jh} = effect of h-th block nested within j-th complete replicate
τ_i = effect der i-th genotype
e_{ijh} = residual plot error associated with y_{ijh}

There are now many different values for the s.e.d., so we just compute the mean s.e.d. as a descriptive measure of accuracy. If we take blocks as fixed (intra-block analysis), the mean s.e.d. is 0.277. If we take the block effect as random (combined inter-intra-block analysis), the mean s.e.d. drops slighly to 0.265. This result is based on a standard REML analysis, which ignores errors in the weights computed from the variance component estimates. If we use thr Kenward-Roger method, which corrects for this bias, the s.e.d. goes up to 0.271. This is still smaller than the s.e.d. for the intra-block analysis, so use of the inter-block information is worthwhile here. The variance component estimates are $\hat{\sigma}_b^2 = 0.06194$ and $\hat{\sigma}^2 = 0.08523$, or in standard deviations, $\hat{\sigma}_b = 0.2488$ and $\hat{\sigma} = 0.2919$.

SAS hints

```
/*blocks fixed*/
ods output diffs=sed_fixed;
proc mixed data=alpha;
class rep block gen;
model y=rep rep*block gen / solution;
lsmeans gen / pdiff;
run;

proc means data=sed_fixed mean;
var StdErr;
run;

/*blocks random*/
ods output diffs=sed_random;
proc mixed data=alpha;
class rep block gen;
model y=rep gen / solution ddfm=KR;
lsmeans gen / pdiff;
random rep*block;
run;

proc means data=sed_random mean;
var StdErr;
run;
```

3.9 Analysis of a resolvable design with checks

Example 6: An augmented design was laid out for 90 entries and 6 checks (**augmented lattice.dat**). The block size was 10. Incomplete blocks were formed according to a 10 × 10 lattice design, in which incomplete blocks can be grouped into complete replicates. Thus, this is a resolvable design, which can be analysed by the same model as the oat data in Section 3.8. Checks are coded as 1001 to 1006, while the 90 entries are coded as 2 to 100 (note that there are no entries with labels 11, 21, 31, 41, 51, 61, 71, 81 and 91). Some of the checks have extra replication. We here consider the trait “yield”.

We may compute the s.e.d. for all pairwise comparisons. Since interest is mainly in the entries, it is useful to only look at the s.e.d. of comparisons involving two entries. When blocks are fixed, the residual variance is estimated as $\hat{\sigma}_e^2 = 1208$ we find for the s.e.d. of comparisons among entries:

minimum 22.4 mean 37.8 maximum 53.9

When blocks are random, the variances are estimated as $\hat{\sigma}_e^2 = 1206$ and $\hat{\sigma}_b^2 = 296$. For the s.e.d. we have

minimum 21.3 mean 36.5 maximum 51.8

Correction for errors in the weights according to Kenward and Roger yields:

minimum 21.6 mean 36.8 maximum 52.4

which is almost the same as without the correction.

3.10 Analysis of a resolvable row-column design

It is often desirable and beneficial to superimpose two blocking structures that are arranged in rows and columns. Such designs are called **row-column designs**. If rows and columns can be grouped to form complete replicates, the design is **resolvable**, as in case of one-way blocking.

Example 7: This example concerns a resolvable row-column design (**rowcol.dat**). A complete replicate is subdivided into incomplete rows and columns. Thus, we have incomplete blocks in both rows and columns. The model now must have effects for rows and columns in place of block effects. We may either take rows and columns as fixed (intra-block analysis) or random (combined intra-block-inter-block analysis).

Table 7: Design and yield data for 35 genotypes of wheat in two replicates with five rows and seven columns (reproduced from Table 4.16 of Kempton and Fox, p. 62)

Design							Yields						
Replicate 1													
20	4	33	28	7	12	30	3.77	3.21	4.55	4.09	5.05	4.19	3.27
10	14	16	21	31	6	18	3.44	4.30	.	3.86	3.26	4.30	3.72
22	11	19	26	29	15	23	3.49	4.20	4.77	2.56	2.87	1.93	2.26
24	25	5	32	2	27	8	3.62	4.52	4.23	3.76	3.61	3.62	4.01
17	9	3	34	13	35	1	3.81	3.75	4.81	3.69	4.61	2.68	4.15
Replicate 2													
31	19	25	34	20	8	6	4.70	7.37	5.03	5.33	5.73	4.70	5.63
24	21	12	4	23	13	3	4.07	5.66	4.98	4.04	4.27	4.10	4.75
11	7	26	5	35	10	30	5.66	6.43	4.59	5.20	4.83	4.70	4.23
33	9	17	18	32	15	2	5.71	6.13	4.63	5.48	5.47	.	4.16
1	27	16	29	14	28	22	5.22	6.16	4.20	4.66	5.54	3.81	3.60

This row-column design is resolvable, because incomplete rows and columns can be grouped into complete replicates. Thus, the model must have an effect for complete replicates, and row and column effects must be nested within replicates. The model is

$$y_{ijhk} = \mu + \gamma_j + r_{jh} + c_{jk} + \tau_i + e_{ijhk}$$

where

y_{ijhk} = yield of i-th genotype in h-th row and k-th column nested within j-th complete replicate
μ = general effect
γ_j = effect of j-th complete replicate
r_{jh} = effect of h-th row within j-th replicate
c_{jk} = effect of k-th column within j-th replicate
τ_i = effect of i-th genotype
e_{ijhk} = residual plot error associated with y_{ijhk}

If we take row and column effects fixed, the residual variance is estimated as $\hat{\sigma}_e^2 = 0.088$, and standard errors of a difference (s.e.d.) show the following minimum, mean, and maximum:

minimum 0.360 mean 0.408 maximum 0.581

If rows and columns are taken as random in order to exploit the inter-block information, we need to estimate three variance components. We find:

$\hat{\sigma}_e^2 = 0.090$ (plots)
$\hat{\sigma}_r^2 = 0.064$ (rows)
$\hat{\sigma}_c^2 = 0.192$ (columns)

The s.e.d. are as follows:

minimum 0.351 mean 0.385 maximum 0.536

This is slightly smaller than for the intra-block analysis. With the Kenward-Roger method we find the following for the s.e.d.:

minimum 0.350 mean 0.403 maximum 0.560

This is still slightly smaller than for the intra-block analysis.

Concluding remark: A special case of a resolvable row-column design is the so-called **lattice square design** where the number of row and of columns are equal and hence the number of treatments must be an exact square. Also, the size of rows and columns equals the square root of the number of treatments. In practice, it is often necessary to have more flexibility regarding block size and number of treatments, so that lattice squares do not play a dominant role.

SAS hints

```
/*rows and columns fixed*/
ods output diffs=rc_sed_fixed;
```

```
proc mixed data=rc;
class rep    row    col    gen;
model y=rep rep*row rep*col gen;
lsmeans gen/diff;
run;

proc means data=rc_sed_fixed mean;
var StdErr;
run;

/*rows and columns random*/
ods output diffs=rc_sed_random;
proc mixed data=rc;
class rep    row    col    gen;
model y=rep gen / ddfm=KR;
random rep*row rep*col;
lsmeans gen/pdiff;
run;

proc means data=rc_sed_random mean;
var StdErr;
run;
```

3.11 Efficiency of resolvable designs

The efficiency of a design depends on the s.e.d or, equivalently, the variance of a difference (v.d.). For the randomized complete block design the variance of a difference is is given by

$$v.d. = \frac{2\sigma_e^2}{r} \quad ,$$

where *r* is the number of replicates and σ_e^2 is the error variance, i.e. the variance of plots within complete blocks (see Section 3.1). By incomplete blocking the error variance can usually be reduced. If, however, the error variance is not reduced by incomplete blocking, then the use of incomplete blocks will result in some loss of efficiency. The reason is that use of incomplete blocks requires making many indirect comparisons among genotypes, while with complete blocks the whole analysis is based on direct comparisons. Indirect comparisons are less accurate than direct comparisons. For example, a direct comparison of two entries in a block has variance $2\sigma_e^2$, because only two observations are involved, while an indirect comparison via a third genotype has variance $4\sigma_e^2$, because four observations are involved.

Example 3 (continued):

		Genotype 1	2	3
Block	2	18	.	33
	3	57	73	.

Direct comparisons:

D_1 = Difference "genotype 1 – genotype 2" = 57 – 73 = –16 (observed in block 3)
D_2 = Difference "genotype 1 – genotype 3" = 18 – 33 = –15 (observed in block 2)

Indirect comparison:

D_3 = Difference "genotype 2 - genotype 3" = Difference "genotype 1 - genotype 3" in block 2
- Difference "genotype 1 - genotype 2" in block 3
$= D_2 - D_1$
$= +1$

		Genotype 1	2	3
Block	2	y_{12}	.	y_{32}
	3	y_{13}	y_{23}	.

$\text{var}(D_1) = \text{var}(y_{13} - y_{23}) = \text{var}(y_{13}) + \text{var}(y_{23}) = 2\sigma_e^2$
$\text{var}(D_2) = \text{var}(y_{12} - y_{32}) = \text{var}(y_{12}) + \text{var}(y_{32}) = 2\sigma_e^2$

$\text{var}(D_3) = \text{var}(y_{12} - y_{32} - (y_{13} - y_{23})) = \text{var}(D_2 - D_1) = \text{var}(D_2) + \text{var}(D_1) = 2\sigma_e^2 + 2\sigma_e^2 = 4\sigma_e^2$

When using incomplete blocks, the s.e.d. (v.d.) usually differs among pairs of genotypes. Thus, it is best to work with the mean v.d. Efficiency of a resolvable design may then be defined as

$$E = \frac{2\sigma_e^2}{r\left[\overline{v.d.}_{IB}\right]}$$

where $\overline{v.d.}_{IB}$ is the mean variance of a difference for the resolvable design based on an combined intra-block-inter-block analysis (random effects for incomplete blocks). Now the error variance will usually be reduced by incomplete blocking. In computing $\overline{v.d.}_{IB}$, we will use the REML estimate of the error variance from an analysis of the resolvable design. Now how should we estimate the larger error variance for the complete block design? A simple approach is to analyse the same data according to a model for a complete block design, regarding complete replicates as complete blocks.

Example 5 (continued): For the α-design analysis of the oats data, we find an average v.d. of

$\overline{v.d.}_{IB} = 0.07337$.

Analysis assuming a complete block design yields

$$v.d. = \frac{2\hat{\sigma}_e^2}{r} = 0.08972 \quad .$$

From this the efficiency is

$$E = \frac{2\hat{\sigma}_e^2}{r\left[\overline{v.d.}_{IB}\right]} = \frac{0.08972}{0.07337} = 1.22$$

Thus, there is a gain in efficiency of about 22%.

For UK cereal trials, Patterson and Hunter (1983) report average efficiency gains by incomplete blocking of 30%.

SAS hints

```
/*analyis according to alpha design - take blocks random for recovery of
  information.*/

ods output diffs=vd_alpha;
proc mixed data=alpha;
class rep block gen;
model y=rep gen / ddfm=KR;
random rep*block;
lsmeans gen/pdiff;
run;

proc print data=vd_alpha;
run;

data vd_alpha;
set vd_alpha;
vd_alpha=StdErr**2;
run;

proc means data=vd_alpha mean;
var vd_alpha;
output out=mean_vd_alpha mean=mean_vd_alpha;
run;

proc print data=mean_vd_alpha;
run;

/*analysis according to RCBD - just drop incomplete block effect from
  model for alpha design analysis*/

ods output diffs=vd_RCBD;
proc mixed data=alpha;
class rep block gen;
model y=rep gen;
lsmeans gen/pdiff;
run;

proc print data=vd_RCBD;
```

```
run;

data vd_RCBD;
set vd_RCBD;
vd_RCBD=StdErr**2;
run;

proc means data=vd_RCBD mean;
var vd_RCBD;
output out=mean_vd_RCBD mean=mean_vd_RCBD;
run;

proc print data=mean_vd_RCBD;
run;

/*merge the two files contining the mean vd for the two analyses
  And then compute efficiency E*/

data efficiency;
merge mean_vd_alpha mean_vd_RCBD;
efficiency=mean_vd_RCBD/mean_vd_alpha;
run;

proc print data=efficiency;
run;
```

3.12 Post blocking

Example 8: A yield trial with 64 oat genotypes was laid out as a 8 × 8 lattice with 3 replicates (H. F. Utz, pers. comm.; **rowcol_fromutz.dat**). The data may be analysed using the model

$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + e_{ijh}$

which was described in Section 3.8. Taking blocks as random for recovery of inter-block information, the variance component estimates are

$\hat{\sigma}_b^2 = 23{,}236$ and
$\hat{\sigma}_e^2 = 12{,}060$,

and the average variance of a difference is 3,206.

Now the layout of incomplete blocks in the field was such that plots can be unequivocally assigned to a particular row and column. These rows and columns can be used for what is called post blocking, i.e., the data are analysed taking rows and columns as additional block factors. In this example, we fit effects for rows and columns nested within replicates. The block and error variances drop to

$\hat{\sigma}_b^2 = 2{,}121$ and
$\hat{\sigma}_e^2 = 6{,}573$.

The variances for rows and columns within replicates are estimated as

$\hat{\sigma}_{rows}^2 = 39{,}172$ and

$\hat{\sigma}^2_{columns} = 12{,}267$.

Thus, rows and columns explain a considerable proportion of the total variation. The average variance of a difference under this model is 2,692 < 3,206. Thus, post-blocking has effected a considerable gain in efficiency.

A note of caution: There are usually several options for post-blocking, depending on the layout. For example, rows and columns may be either nested within replicates, or one may fit effects for rows and columns which extend across replicates. There being several options, one may be inclined to try them all and settle for the one yielding the smallest s.e.d. This is not a good idea, however, because it will tend to yield too small values of s.e.d. Clearly, the more options there are for post-blocking, the more options there are to exploit chance variation and pick a model which seems to be better than others purely by chance.

To give an extreme analogous example, assume you want to prove that a die is biased towards sixes. You toss the die five times and count the number of times a six faced up. After the first five tosses you find one six. This does not seem to prove your claim. So you repeat and find only two sixes in five tosses. This still does not seem to prove your claim. So you keep repeating the experiment until you find five sixes in a row and then say, "Look, didn't I tell you the die was manipulated?". --Fitting a large battery of models without solid scientific background and then selecting the one giving the most favorable result is a bit like tossing the die until you hit five sixes in a row.

The best insurance against this type of data-dredging is keep to the rule "**analyse-as-randomize**". The second-best insurance is to decide on only one way or just a few ways of post-blocking **before** completing the experiment ("pre-experiment") and to have good reasons for expecting that the contemplated types of post-blocking will be effective. For example, one may anticipate edge effects, i.e., plots at the margin of the experimental area are expected to yield differently than plots inside the experimental layout. Tractor wheeling may cause substantial differences among rows parallel to the wheel tracks, so using rows as block effects may be useful. It is usually worth thinking twice, however, whether post-blocking can be turned into "pre-blocking". If one type of post-blocking seems to be preferable from the start, e.g., because one anticipates certain soil trends, it may be possible to pre-block accordingly, i.e., to find a design that is optimal for the blocking structure expected to best capture trends.

Post-blocking is sometimes used "post-mortem", e.g. when a disease has been found to have affected many plots. Then, affected plots can be regarded as one block, while unaffected plots form a second blocks. Post-blocking may be quite effective in such cases, and it will reduce the error variance estimates, but it should be kept in mind that it is bound to yield somewhat too optimistic s.e.d., because the model has been selected after seeing the data.

The issue of post-blocking is quite controversial, even among statisticians (Gilmour, 2000; Williams and Fu, 1999), and there is no consensus as to the best strategy. My personal view is that post-blocking does have its role to play in the analysis of individual breeding trials, e.g., as a tool to save a trial post-mortem in case of unforeseen events such as pest infestation, or foreseen events such as tractor-wheeling and edge effects, but it should be used with some caution. Options for post-blocking should be specified pre-experiment, and the number of options should be limited.

4 Sub-sampling

4.1 The sorghum experiment

Example 9: In a greenhouse experiment with sorghum, three intensities of double use of grain and leaves were tested: (1) control (grain only); (2) removal of all leaves except top leaf; (3) removal of all leaves except top six leaves (Piepho, 1997). The leaves were removed at four different dates within three weeks. The design was a randomized complete block design with four replicates. The plants were planted individually into pots (one plant per pot). One replicate consisted of 10 plants (pots) per treatment, which were placed on a tray. Thus, a complete replicate was made up of three trays. Measurements were taken on individual plants. Here, we consider analysis of thousand kernel weight (g) (see Table 8). Some plants died before the first leaves were harvested. The missing data pattern therefore is independent of treatments, which is an important requirement for the type of analysis we are considering. In statistical terms, the data meet the **missing completely at random (MCAR)** assumption. Because of the missing observations, the data is unbalanced. Unbalancedness should be accounted for a fully efficient analysis. As we shall see, optimal analysis will involve weighting.

Table 8: Thousand kernel weight of sorghum in greenhouse experiment on double use (fodder and grain production; **sorghum.dat**).

		Running number of pot / plant									
Treatment	Block	1	2	3	4	5	6	7	8	9	10
1	1	41.89	29.97	27.82	29.68	27.84	33.93	34.77	30.00	27.71	32.66
	2	28.21	34.80	31.25	34.35	37.50	36.74	32.23	38.35	29.50	
	3	41.54	44.19	40.44	35.20	26.76	32.37	31.18	34.23	34.22	25.01
	4	35.26	26.65	35.47	29.84	34.38	26.79	41.39	40.60	28.16	28.11
2	1	41.28	32.86	35.15		21.76	34.60	19.36	46.22	41.64	32.91
	2	27.55	40.89	33.14	29.62		48.43	25.39	35.45	27.73	29.45
	3	32.27	38.62	24.18	29.34	22.77		23.87	28.01	24.70	23.70
	4	42.36	35.61	17.47	23.55	18.21	33.21	32.76	26.47	21.68	
3	1	27.40	32.66	24.15	27.98	35.43		31.94	32.24	29.87	
	2	44.37	30.01		36.49	40.40	29.48	24.69	35.03	27.48	
	3	35.98	32.82	27.16	29.81	33.91	28.28	29.67	33.66		29.50
	4	36.58	27.64	27.53	26.23		31.00	31.36	29.67	30.12	27.33

Objective: ANOVA and mean comparisons.

Modelling: Fig. 3 shows a sketch of a complete block.

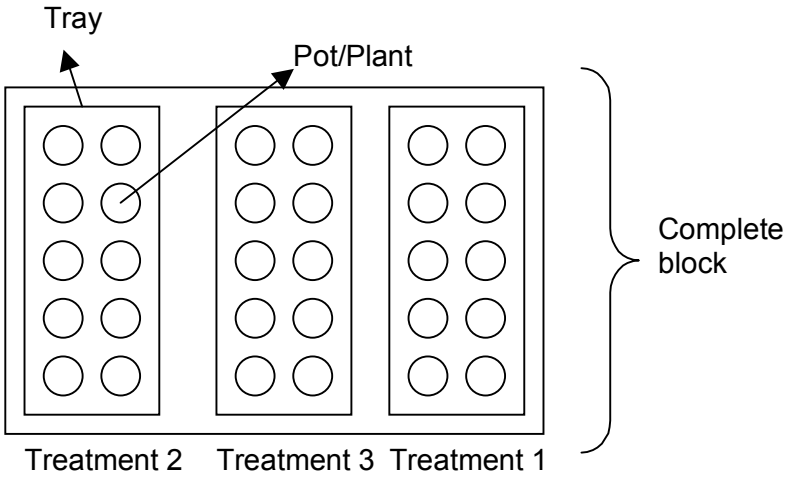


Fig. 3a: Sketch of design for one block (out of four!) of sorghum experiment.

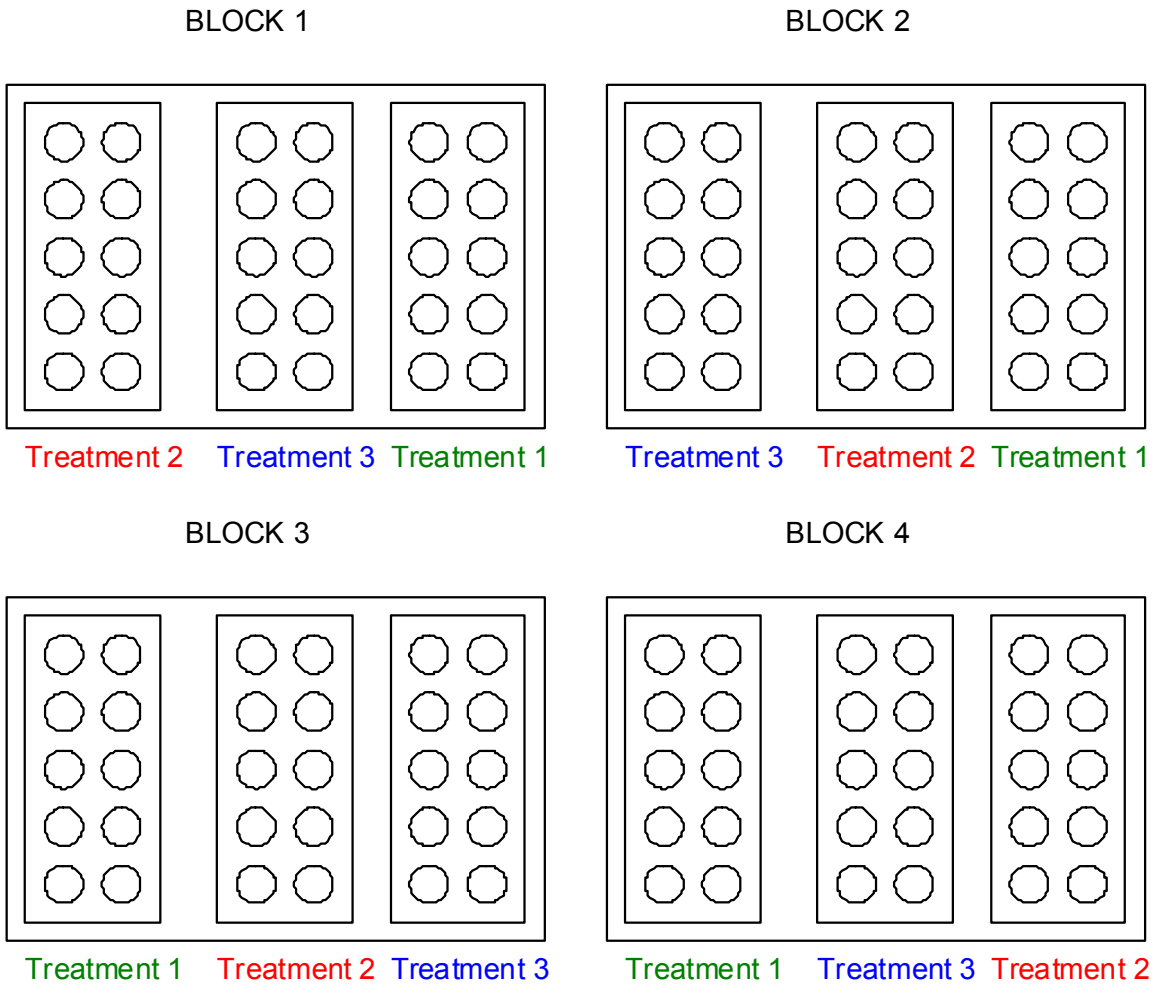


Fig. 3b: The whole experiment with four complete blocks.

The experiment involves sub-sampling with pots as sampling units. The trays are the actual randomization units and so the number of „true“ replications is the number of trays per treatment, while plants are merely repeated measurements on the same unit. This structure

needs to be accounted for at the modelling stage. In particular there need to be two random error terms, one corresponding to trays (between trays) and one corresponding to plants (within trays). A linear model for this data structure is as follows:

$y_{ijk} = \mu + \tau_i + b_j + u_{ij} + e_{ijk}$

where

y_{ijk} = measurement of k-th plant of i-th treatment and j-th block
μ = general mean
τ_i = effect of i-th treatment
b_j = effect of j-th blocks
u_{ij} = effect of ij-th tray (random)
e_{ijk} = effect of ijk-th plant (random residual)

So the model has two random effects, i.e., the tray effect u_{ij} and the plant effect e_{ijk}. We make the usual distributional assumptions

$u_{ij} \sim N(0, \sigma_u^2)$ and
$e_{ijk} \sim N(0, \sigma_e^2)$,

where N(0, .) denotes the normal distribution with zero mean and given variance.

Analysis: One could consider a **two-stage analysis**, by which tray means are computed in the first step ($\bar{y}_{ij\bullet}$), which are then subjected to an ANOVA for a block design at the second stage. In fact, this analysis will yield reasonable results when the data are not too unbalanced. The potential problem with this approach is that means do not have a constant variance under the assumed model when some observations are missing. We then have:

$$\text{var}(\bar{y}_{ij\bullet}) = \sigma_u^2 + \frac{\sigma_e^2}{r_{ij}} \quad ,$$

where r_{ij} is the number of plants on the ij-th tray. Thus, the variance of a tray mean depends on the sample size per tray (r_{ij}), and so variances are heterogeneous whenever r_{ij} is not constant across trays. A simple ANOVA will thus violate the underlying assumptions.

The mixed model for the plant data ($y_{ijk} = \mu + \tau_i + \beta_j + u_{ij} + e_{ijk}$) can account for this heterogeneity by appropriate weighting. Analysis of plant data proceeds in a single stage. The optimal weight given to a tray mean will be inversely proportional to its variance. The general method of estimation is denoted as weighted least squares, and when weights are computed from the variances in a particular form, the resulting estimator will have the BLUE (best linear unbiased estimator) property.

For the case at hand the weighted estimator does not have a simple algebraic form. If analysis is by a mixed model package with the appropriate model, the optimal weighting will be obtained in an automatic fashion. We will use this example to show the general form of mixed model that is used by a mixed model package and to illustrate general methods for inferences on fixed effects, which include treatment comparisons.

We fit the model for plant data (not tray means!) using a mixed model package.

Output:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	2	6.15	1.43	0.3097
block	3	6.13	1.11	0.4136

Least Squares Means

Effect	treat	Estimate	Standard Error
treat	1	33.1393	1.0965
treat	2	30.6169	1.1288
treat	3	31.1691	1.1546

Differences of Least Squares Means

Effect	treat	_treat	Estimate	Standard Error	DF	t Value	Pr > |t|
treat	1	2	2.5223	1.5737	5.87	1.60	0.1611
treat	1	3	1.9702	1.5922	6.14	1.24	0.2612
treat	2	3	-0.5521	1.6147	6.5	-0.34	0.7432

The so-called Wald-F-value for treatments is 1.43 and is not significant. Also, mean comparisons are not significant. Note that standard errors of means and of differences are not constant. This is a result of unbalancedness. The degrees of freedom are not integer here, because we have used the Kenward-Roger method for approximating the degrees of freedom.

SAS hints

```
data s;
input
treat    block    plant     TKG  ;
tray=100*block+treat;
datalines;
  1        1         1     41.89
  1        1         2     29.97
 <more data>
  3        4         8     29.67
  3        4         9     30.12
  3        4        10     27.33
;
/*option 1*/
```

```
proc mixed data=s;
class treat block;
model TKG=treat block / ddfm=KR;
random treat*block; /*tray effect*/
lsmeans treat/pdiff;
run;

/*option 2*/
proc mixed data=s;
class treat block tray;
model TKG=treat block / ddfm=KR;
random tray; /*tray effect*/
lsmeans treat/pdiff;
run;
```

***4.2 General results for inference on fixed effects and application to examples**

Any linear mixed model can be written in matrix form as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \ ,$$

where $\boldsymbol{y}$ is a data vector with n observations, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices, $\boldsymbol{\beta}$ is a parameter vector of fixed effects, $\boldsymbol{u}$ is a vector of random effects, and $\boldsymbol{e}$ is a vector of errors. In a linear model with fixed effects, errors in $\boldsymbol{e}$ are independent and normally distributed with zero mean ($\mathbf{0}$) and variance σ^2, which may be expressed as

$$\boldsymbol{e} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}\sigma^2) \ ,$$

where MVN denotes the multivariate normal distribution. In a linear mixed model this assumption may be generalized as

$$\boldsymbol{e} \sim \text{MVN}(\mathbf{0}, \boldsymbol{R}) \ ,$$

where $\boldsymbol{R}$ is a variance-covariance-matrix with variances on the diagonal and covariances on the off-diagonals. Moreover all random effects in $\boldsymbol{u}$ are normal with zero mean ($\mathbf{0}$) and variance-covariance matrix $\boldsymbol{G}$:

$$\boldsymbol{u} \sim \text{MVN}(\mathbf{0}, \boldsymbol{G}) \ .$$

With these specifications, the distribution of observed data becomes

$$\boldsymbol{y} \sim \text{MVN}(\boldsymbol{X\beta}, \boldsymbol{V}) \ ,$$

where

$$\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R} \ .$$

Example 2 (continued): In Section 3.1 we considered the following unbalanced dataset.

		Genotype		
		1	2	3
Block	1	12	18	37
	2	18	32	33
	3	57	73	.

This may be analysed according to the model

$$y_{ij} = \mu + b_j + \tau_i + e_{ij} \quad ,$$

where y_{ij} is the observed yield of the i-th genotype in the j-th block, b_j is the effect of the j-th block, τ_i is the effect of the i-th genotype, and e_{ij} is the residual effect corresponding to y_{ij}. If block effects are considered as random, this becomes a mixed model, which can be expressed in matrix form as

$$\begin{pmatrix} 12 \\ 18 \\ 57 \\ 18 \\ 32 \\ 73 \\ 37 \\ 33 \end{pmatrix} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{\beta} + \boldsymbol{Z} \times \boldsymbol{u} + \boldsymbol{e}$$

$$\text{var}(\boldsymbol{u}) = \boldsymbol{G} = \text{var}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} \sigma_b^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_b^2 \end{pmatrix}$$

$$\text{var}(\boldsymbol{e}) = \boldsymbol{R} = \text{var}\begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{31} \\ e_{32} \end{pmatrix} = \begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 \end{pmatrix}$$

$$\text{var}(\boldsymbol{y}) = \boldsymbol{V} = \text{var}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} \sigma_b^2+\sigma_e^2 & 0 & 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2 & 0 \\ 0 & \sigma_b^2+\sigma_e^2 & 0 & 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2 \\ 0 & 0 & \sigma_b^2+\sigma_e^2 & 0 & 0 & \sigma_b^2 & 0 & 0 \\ \sigma_b^2 & 0 & 0 & \sigma_b^2+\sigma_e^2 & 0 & 0 & \sigma_b^2 & 0 \\ 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2+\sigma_e^2 & 0 & 0 & \sigma_b^2 \\ 0 & 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2+\sigma_e^2 & 0 & 0 \\ \sigma_b^2 & 0 & 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2+\sigma_e^2 & 0 \\ 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2 & 0 & 0 & \sigma_b^2+\sigma_e^2 \end{pmatrix}$$

In this matrix formulation, the variances of elements of a random vector are on the diagonal, while the covariances are on the off-diagonal. For example, the covariance between y_{11} and y_{21}, i.e., of two yields in the same block, is

$$\text{cov}(y_{11}, y_{21}) = \sigma_b^2 \quad .$$

The covariance equals the block variance, because these two observations share the same random block effect.

Example 9 (continued): For the sorghum experiment the model can be written as

$$\begin{pmatrix} y_{111} \\ y_{112} \\ \cdot \\ \cdot \\ y_{11r_{11}} \\ y_{121} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ y_{34r_{34}} \end{pmatrix} = \begin{pmatrix} 1&1&0&0&1&0&0&0 \\ 1&1&0&0&1&0&0&0 \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot \\ 1&1&0&0&1&0&0&0 \\ 1&1&0&0&0&1&0&0 \\ 1&\cdot& & & & & &\cdot \\ 1& &\cdot& & & & &\cdot \\ 1& & &\cdot& & & &\cdot \\ 1& & & &\cdot& & &\cdot \\ 1& & & & &\cdot& &\cdot \\ 1&0&0&1&0&0&0&1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} + \begin{pmatrix} 1&0&0&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0&0&0 \\ \cdot& &\cdot& & & & & & & & &\cdot \\ \cdot& & &\cdot& & & & & & & &\cdot \\ 1&0&0&0&\cdot&0&0&0&0&0&0&0 \\ 0&1&0&0&0&\cdot&0&0&0&0&0&0 \\ \cdot& & & & & &\cdot& & & & &\cdot \\ \cdot& & & & & & &\cdot& & & &\cdot \\ \cdot& & & & & & & &\cdot& & &\cdot \\ \cdot& & & & & & & & &\cdot& &\cdot \\ \cdot& & & & & & & & & &\cdot&\cdot \\ 0&0&0&0&0&0&0&0&0&0&0&1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{21} \\ u_{22} \\ u_{23} \\ u_{23} \\ u_{31} \\ u_{32} \\ u_{33} \\ u_{34} \end{pmatrix} + \begin{pmatrix} e_{111} \\ e_{112} \\ \cdot \\ \cdot \\ e_{11r_{11}} \\ e_{121} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ e_{34r_{34}} \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X} \times \boldsymbol{\beta} + \boldsymbol{Z} \times \boldsymbol{u} + \boldsymbol{e}$$

When fitting a fixed effects linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$, the objective function to be minimizes is the error sum of squares given by $SS_{error} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$. In a mixed model, this is replaced by the weighted error sum of squares $SS_{error} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$. This objective

function results from an optimization of the Likelihood function under the assumed multivariate normal distribution for $\boldsymbol{y}$, which is given by

$$L(\boldsymbol{\beta},\boldsymbol{V} \mid \boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\boldsymbol{V}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right] . \qquad \text{(x)}$$

Taking the logarithm, the log-likelihood is (ignoring constants)

$$\log L = -\frac{1}{2}\log(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) .$$

The maximum with respect to $\boldsymbol{\beta}$ must maximize $\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$, which is the weighted error sum of squares. When variance components in $\boldsymbol{V}$ are known, the resulting weighted least squasres solution is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} ,$$

where $\boldsymbol{M}^{-}$ is a so-called *g*-inverse of $\boldsymbol{M}$. For comparison, the ordinary least squares solution under a fixed effects model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ is

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

The only difference is the use of the variance-covariance matrix $\boldsymbol{V}$ for weighting. When data are uncorrelated, use of $\boldsymbol{V}$ has the effect of giving observations with small variance (i.e. larger precision) a larger weight. In practice, $\boldsymbol{V}$ is usually unknown and needs to be estimated. As indicated in Section 3.5, most packages use either ML or REML. Briefly, ML maximizes the full likelihood (log L) with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{V}$, while REML considers the restricted likelihood for contrasts

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{y} ,$$

where $\boldsymbol{H}$ is chosen such that $\boldsymbol{z}$ is free of the fixed effects $\boldsymbol{\beta}$, i.e., $E(\boldsymbol{z}) = \boldsymbol{0}$. The restricted (or residual) likelihood then is

$$L_R(\boldsymbol{V} \mid \boldsymbol{z} = \boldsymbol{H}\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}'|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{H}\boldsymbol{y})'(\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}')^{-1}(\boldsymbol{H}\boldsymbol{y})\right] .$$

Taking the logarithm, the restricted log-likelihood is (again ignoring constants)

$$\log L_R = -\frac{1}{2}\log(\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}') - \frac{1}{2}(\boldsymbol{H}\boldsymbol{y})'(\boldsymbol{H}\boldsymbol{V}\boldsymbol{H}')^{-1}(\boldsymbol{H}\boldsymbol{y}) .$$

The key point here is that the restricted log-likelihood (log L_R) does not depend on the fixed effects, and it accounts for the degrees of freedom for fixed effects in the model. For

computational details regarding the maximization of log L or log L_R see, e.g., Searle et al. (1992) and McCulloch and Searle (2001).

Now consider tests of hypotheses regarding fixed effects. In general form these can be written

$$H_0 : \boldsymbol{K\beta} = 0 \ ,$$

where $\boldsymbol{K}$ is a matrix of suitably chosen constants.

Example 9 (continued): In the sorghum experiment the **global null hypothesis** of no treatment differences is

$$H_0 : \tau_1 = \tau_2 = \tau_3 \ .$$

This can also be represented by two simple null hypotheses, e.g.

$$H_{01} : \tau_1 = \tau_2 \ \text{ and } \ H_{02} : \tau_1 = \tau_3 \, .$$

This has the following equivalent representation:

$$H_{01} : \tau_1 - \tau_2 = 0 \ \text{ and } \ H_{02} : \tau_1 - \tau_3 = 0 \, .$$

Note that the two simple null hypotheses also imply that $H_{03} : \tau_2 - \tau_3 = 0$ holds, which in turn implies $H_0 : \tau_1 = \tau_2 = \tau_3$. In matrix form, the two simple null hypotheses H_{01} and H_{02} can be stated as

$$\mathrm{H}_0 : \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\boldsymbol{K} \qquad\qquad \times \ \boldsymbol{\beta} \ = \ \mathbf{0}$$

For testing $H_0 : \boldsymbol{K\beta} = 0$, one can compute the Wald-statistic

$$\chi^2 = \hat{\boldsymbol{\beta}}' \boldsymbol{K}' \left(\boldsymbol{K} \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X} \right)^{-} \boldsymbol{K}' \right)^{-1} \boldsymbol{K} \hat{\boldsymbol{\beta}} \ .$$

When $\boldsymbol{V}$ is known, this has an exact χ^2-distribution with degrees of reedom equal to rank($\boldsymbol{K}$), which is the number of independent rows in $\boldsymbol{K}$. For example, when testing for the equality of t

treatments, we have $\text{rank}(\boldsymbol{K}) = t - 1$, which is the usual treatment degrees of freedom. If variances in $\boldsymbol{V}$ need to be estimated, it is better to use theWald-type F-statistic

$$F = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{K}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\hat{\boldsymbol{\beta}}}{\text{rank}(\boldsymbol{K})} ,$$

where $\hat{\boldsymbol{V}}$ is the REML estimator of $\boldsymbol{V}$. This can be refered to an F-distribution with $\text{rank}(\boldsymbol{K})$ numerator degrees of freedom and denominator degrees of freedom approximated by one of several methods. The preferred method for simple mixed models is the Kenward-Roger (1997) method, a form of Satterthwaite method with a further adjustment accounting for estimation errors in the variance components.

When $\text{rank}(\boldsymbol{K}) = 1$, i.e. we are testing a simple hypothesis such that $\boldsymbol{K}$ has just a single row, $\sqrt{F}$ is equivalent to a t-statistic

$$t = \frac{\boldsymbol{K}\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{K}\left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{K}'}} .$$

It is important to note that the Wald-type F-statistic recovers the standard F-test in fixed effects linear models of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. For fixed effects models we have

$$\boldsymbol{V} = \boldsymbol{I}\sigma^2 ,$$

and hence the F-statistic reduces to

$$F = \frac{\hat{\boldsymbol{\beta}}'\boldsymbol{K}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\hat{\boldsymbol{\beta}}}{s^2\text{rank}(\boldsymbol{K})} ,$$

where

$$s^2 = \frac{\boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\right)\boldsymbol{y}}{n - \text{rank}(\boldsymbol{X})}$$

is the estimator of the residual variance.

Example 10: This may be illustrated using a simple one-way ANOVA, which may be based on the scalar model

$$y_{ij} = \mu_i + e_{ij} ,$$

where μ_i is the mean of the i-th treatment ($i = 1, .., t$) and $e_{ij} \sim N\left(0, \sigma^2\right)$. For three treatments and four replicates the design matrix and parameter vector are

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \ \boldsymbol{K} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

More generally for the one-way ANOVA with t treatments, $\boldsymbol{K} = \left(\mathbf{1}_{t-1} \quad -\boldsymbol{I}_{t-1}\right)$, where $\mathbf{1}_{t-1}$ is a vector of ones and $\boldsymbol{I}_{t-1}$ is an $(t-1)\times(t-1)$ identity matrix. Moreover, $\boldsymbol{X}'\boldsymbol{X} = r\boldsymbol{I}_{t-1}$, where r is the number of replicates per treatment. Furthermore, the following results hold:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\boldsymbol{y} = r^{-1}\boldsymbol{I}\begin{pmatrix} y_{1\bullet} \\ y_{2\bullet} \\ y_{3\bullet} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1\bullet} \\ \bar{y}_{2\bullet} \\ \bar{y}_{3\bullet} \end{pmatrix} = \bar{\boldsymbol{y}}_{\bullet}, \text{ where } \bar{\boldsymbol{y}}_{\bullet} \text{ denotes the vector of treatment means,}$$

$$\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}' = r^{-1}\boldsymbol{K}\boldsymbol{K}' = r^{-1}\left(\mathbf{1} \quad -\boldsymbol{I}_{t-1}\right)\begin{pmatrix} \mathbf{1}'_{t-1} \\ -\boldsymbol{I}_{t-1} \end{pmatrix} = r^{-1}\left(\boldsymbol{I}_{t-1} + \boldsymbol{J}_{t-1}\right),$$

where $\boldsymbol{J}_{t-1} = \mathbf{1}_{t-1}\mathbf{1}'_{t-1}$ is a matrix $(t-1)\times(t-1)$ matrix of ones everywhere,

$$\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1} = r\left(\boldsymbol{I}_{t-1} - t^{-1}\boldsymbol{J}_{t-1}\right)$$

$$\left(\boldsymbol{K}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\left[\boldsymbol{K}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\right]\hat{\boldsymbol{\beta}}$$

$$\boldsymbol{K}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K} = r\begin{pmatrix}\mathbf{1}'_{t-1}\\ -\boldsymbol{I}_{t-1}\end{pmatrix}\left(\boldsymbol{I}_{t-1}-t^{-1}\boldsymbol{J}_{t-1}\right)\left(\mathbf{1}_{t-1} \quad -\boldsymbol{I}_{t-1}\right)$$

$$= r\begin{bmatrix}\mathbf{1}'_{t-1}\left(\boldsymbol{I}_{t-1}-t^{-1}\boldsymbol{J}_{t-1}\right)\mathbf{1}_{t-1} & -\mathbf{1}'_{t-1}\left(\boldsymbol{I}_{t-1}-t^{-1}\boldsymbol{J}_{t-1}\right)\\ -\left(\boldsymbol{I}_{t-1}-t^{-1}\boldsymbol{J}_{t-1}\right)\mathbf{1}_{t-1} & \left(\boldsymbol{I}_{t-1}-t^{-1}\boldsymbol{J}_{t-1}\right)\end{bmatrix} = r\left(\boldsymbol{I}_{t}-t^{-1}\boldsymbol{J}_{t}\right)$$

and hence the numerator of the Wald-type F-statistic becomes

$$\hat{\boldsymbol{\beta}}'\left[\boldsymbol{K}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\right]\hat{\boldsymbol{\beta}} = r\bar{\boldsymbol{y}}'_{\bullet}\left(\boldsymbol{I}_{t}-t^{-1}\boldsymbol{J}_{t}\right)\bar{\boldsymbol{y}}_{\bullet} = r\sum_{i=1}^{t}\left(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet}\right)^{2} \ ,$$

which is just the usual treatment sum of squares for the one-way ANOVA. Also, it can be similarly shown that

$$s^{2} = \frac{\boldsymbol{y}'\left(\boldsymbol{I}-\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\right)\boldsymbol{y}}{n-\operatorname{rank}\left(\boldsymbol{X}\right)} = \frac{\sum_{i,j}\left(y_{ij}-\bar{y}_{i\bullet}\right)^{2}}{n-t},$$

such that

$$F = \frac{\boldsymbol{K}'\hat{\boldsymbol{\beta}}'\left(\boldsymbol{K}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{K}'\right)^{-1}\boldsymbol{K}\hat{\boldsymbol{\beta}}}{s^{2}\operatorname{rank}\left(\boldsymbol{K}\right)} = \frac{n-t}{t-1}\frac{r\sum_{i=1}^{t}\left(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet}\right)^{2}}{\sum_{i,j}\left(y_{ij}-\bar{y}_{i\bullet}\right)^{2}},$$

which is seen to be the usual one-way ANOVA F-statistic. The example shows that the Wald-type F-test for a mixed model recovers the well-known simple F-test in case of a one-way ANOVA.

Kenward-Roger method

The Kenward-Roger method does two things:

(1) Adjust the denominator degrees of freedom, given the estimates of the variance components and their associated asymptotic variance-covariance matrix (see chapter on split-plot designs)

(2) Correct the estimate of the standard errors of adjusted means and their differences (this has an effect only when adjusted means are NOT equal to simple averages!!). This addresses the problem with the errors of the weights when using the generalized least squares estimator $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}$. But this problem vanishes if this estimator happens to coincide with the ordinary least squares estimator $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\boldsymbol{y}$, which does not require any weights. This is the case when the data is balanced.

My recommendation is to always use the Kenward-Roger method with mixed models; you can't go wrong with this.

When the variance-covariance matrix is non-linear in the parameters, such as with spatial and repeated-measures models (Chapters 6 and 7), it is better to use a second-order extension of the Kenward-Roger method, available via the option ddfm=KENWARDROGER2.

5 Split-plots

5.1 Randomisation

Example 11: A field trial was conducted to test six levels of nitrogen fertilizer (N1, N2, N3, N4, N5, N6) and four rice varieties (V1, V2, V3, V4) (**rice.dat**). The trial was laid out in split plots with three replicates. Randomization was as follows:

Step 1: The field was divided into three blocks (replicates). Each block was divided into six **main plots**. For every block separately, the six fertilizer treatments were randomly allocated to main plots.

N1	N2	N4	N6	N3	N5

Block I

N3	N2	N5	N1	N6	N4

Block II

N4	N6	N3	N1	N2	N5

Block III

Step 2: Every main plot was split into four sub plots to accommodate the four varieties. Separately for each main plot, the varieties were randomly allocated to the four sub-plots.

N1V3	N2V4	N4V1	N6V3	N3V2	N5V1
N1V2	N2V1	N4V4	N6V1	N3V1	N5V4
N1V4	N2V3	N4V2	N6V4	N3V3	N5V2
N1V1	N2V2	N4V3	N6V2	N3V4	N5V3

Block I

N3V2	N2V2	N5V2	N1V4	N6V3	N4V1
N3V1	N2V4	N5V3	N1V3	N6V1	N4V3
N3V4	N2V3	N5V4	N1V1	N6V4	N4V4
N3V3	N2V1	N5V1	N1V2	N6V2	N4V2

Block II

N4V1	N6V1	N3V4	N1V1	N2V1	N5V3
N4V4	N6V2	N3V3	N1V3	N2V4	N5V1
N4V2	N6V4	N3V2	N1V4	N2V2	N5V4
N4V3	N6V3	N3V1	N1V2	N2V3	N5V2

Block III

It is important to recognize that nitrogen, the so-called **main plot factor**, was randomized according to a randomized complete block design. Varieties, corresponding to the so-called **sub-plot factor**, were also randomized according to a randomized complete block design, taking main plots as block. This type of split-plot design is the most common form, but there are many other forms of „the" split-plot design depending on the design according to which the main plot and sub plot factors are randomized.

5.2 Basic modelling

The standard analysis is a two-way ANOVA (variety × fertilizer), followed by multiple comparison of means. Split-plot designs have two randomization units, which need to be represented in the linear model: main plots and sub plots. As a general principle, each randomization units needs to be represented by a random effect, so each randomization unit has its own error term. In the example, fertilizer levels are compared at the main plot level, so the **main plot error** is the relevant error term. The varieties are compared at the sub plot level, and so the **sub plot error** is the relevant error term.

In the present example the model needs to contain a block effect because the main plots were randomized in complete blocks. In addition, a two-way model for the treatment factors is needed. Let y_{ijh} denote the observed response for the i-th nitrogen level (N) and the j-th variety (V) in block h.

Blocks: $h = 1, ..., r$
Nitrogen (N): $i = 1, ..., a$ (main plot factor)
Variety (V): $j = 1, ..., b$ (sub plot factor)

The expected value of an observation can be written as

$$E(y_{ijh}) = \eta_{ijh} = \mu + b_h + \tau_{ij} \quad ,$$

where

μ = general effect (intercept)
b_h = effect of h-th block
τ_{ij} = effect of ij-th treatment

The treatment effect τ_{ij} can be factorized as usual:

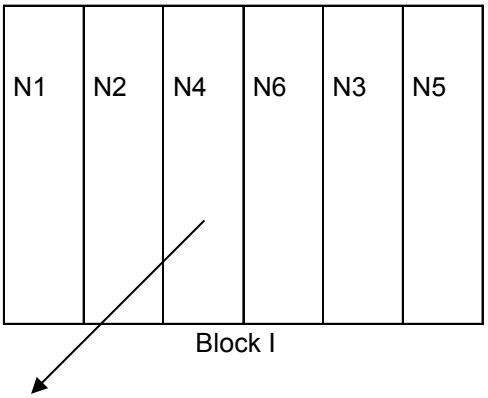
$\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$

where

α_i = main effect of i-th N-level
β_j = main effect of j-th variety
$(\alpha\beta)_{ij}$ = interaction

This factorization is best used when both treatment factors are qualitative. When a treatment factor is quantitative, some form of regression model may be used.

Now consider the error terms required fort his design, i.e., the main plot error and the sub plot error. Every combination of block and N-level corresponds to a main plot.



Main plot for $i = 4$ (4-th N-level) and
$h = 1$ (1st block)

Thus, every combination of block × N-level must have its own error term. This is achieved by adding an effect f indexed by block and N-level:

f_{ih} = main plot error
= error of main plot for block h with N-level i

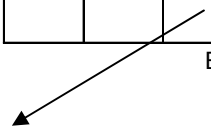
It is assumed that main plot errors f_{ih} are normal with zero mean and variance σ_f^2:

$f_{ih} \sim N(0, \sigma_f^2)$

Similarly, every combination of block × N-level × variety corresponds to a sub plot.

N1V3	N2V4	N4V1	N6V3	N3V2	N5V1
N1V2	N2V1	N4V4	N6V1	N3V1	N5V4
N1V4	N2V3	N4V2	N6V4	N3V3	N5V2
N1V1	N2V2	N4V3	N6V2	N3V4	N5V3

Block I



Sub plot for $i = 4$ (4-th N-level),
$j = 3$ (3-th variety), and
$h = 1$ (1st block) □

So the sub plot error is indexed by block, N-level and variety:

e_{ijh} = sub plot error
= error effect of sub plot in block h with i-th N-level and j-th variety

The sub plot error is assumed to be normal with zero mean and variance σ^2:

$e_{ijh} \sim N(0, \sigma^2)$.

The complete model is

$$y_{ijh} = E(y_{ijh}) + f_{ih} + e_{ijh}$$
$$= \eta_{ijh} + f_{ih} + e_{ijh}$$
$$= \mu + b_h + \tau_{ij} + f_{ih} + e_{ijh} ,$$

and using the factorial model for τ_{ij}:

$$y_{ijh} = \mu + b_h + \alpha_i + \beta_j + (\alpha\beta)_{ij} + f_{ih} + e_{ijh} .$$

Analysis: A peculiarity of the split-plot design is that two error terms are involved, and this affects F-tests as well as multiple comparisons of means. We might just observe that the above linear model is a mixed model, and that general methods such as those outlined in Section 4.2 are available for producing adequate t-tests and F-tests, and say no further. It is perhaps instructive, however, to relate these more recent techniques to the classical ANOVA for a split plot design (Cochran and Cox, 1957). In case we have complete data, the split-plot ANOVA has the following format:

Source	d.f.	Sum of squares	Mean square
Main plot stratum:			
Blocks	$(r-1)$	SS_{blocks}	
Main effect A (main plot factor)	$(a-1)$	SS_A	
Error (A) (main plot error)	$(a-1)(r-1)$	$SS_{error(A)}$	E_a
Sub plot stratum:			
Main effect B (sub plot factor)	$(b-1)$	SS_B	
Interaction	$(a-1)(b-1)$	SS_{AB}	
Error (B) (sub plot error)	$a(b-1)(r-1)$	$SS_{error(B)}$	E_b

r = no. of blocks
a = no. of levels for factor A
b = no. of levels for factor B
E_a and E_b are abbreviations for the mean squares for main plot and sub plot error; SS with different subscripts denote the various sums of squares.

We can compute marginal means of the main plot factor (nitrogen in our example). These means can be thought of as computed in two steps. In the first step, we compute means per main plot. In the second step we compute means for N-levels. This corresponds to a standard ANOVA of main plot means, and the error term for this analysis has $(a-1)(r-1)$ degrees of freedom, which is seen to be the same as the main plot error d.f. in the ANOVA table given above. By contrast, sub plot main effect and interaction are tested against a different error term, the sub plot error.

The need to use different error terms can be justified by looking at the expected mean squares of an ANOVA table for the split plot design:

Table 9: Expected mean squares in the split-plot ANOVA, assuming balanced data.

Source	d.f.	Expected *MS* - E(*MS*)	H_0 tested by *MS*	Expected *MS* under H_0
Main plot stratum:				
Blocks	$(r-1)$	$\sigma^2 + b\sigma_f^2 + \frac{ab}{r-1}\sum_{h=1}^{r}\left(b_h - \bar{b}_\bullet\right)^2$	$b_1 = b_2 = \ldots$	$\sigma^2 + b\sigma_f^2$
Main effect A	$(a-1)$	$\sigma^2 + b\sigma_f^2 + \frac{br}{a-1}\sum_{i=1}^{a}\left(\alpha_i - \bar{\alpha}_\bullet\right)^2$ (*)	$\alpha_1 = \alpha_2 = \ldots$	$\sigma^2 + b\sigma_f^2$
Error (A)	$(a-1)(r-1)$	$\sigma^2 + b\sigma_f^2$	None	$\sigma^2 + b\sigma_f^2$
Sub plot stratum:				
Main effect B	$(b-1)$	$\sigma^2 + \frac{ar}{(b-1)}\sum_{j=1}^{b}\left(\beta_j - \bar{\beta}_\bullet\right)^2$ (*)	$\beta_1 = \beta_2 = \ldots$	σ^2
Interaction	$(a-1)(b-1)$	$\sigma^2 + \frac{r}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\left[(\alpha\beta)_{ij} - \overline{(\alpha\beta)}_{i\bullet} - \overline{(\alpha\beta)}_{\bullet j} + \overline{(\alpha\beta)}_{\bullet\bullet}\right]^2$	$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots$	σ^2
Error (B)	$a(b-1)(r-1)$	σ^2	None	σ^2

(*) Assuming there is no interaction!

r = no. of blocks
a = no. of levels for factor A
b = no. of levels for factor B

The appropriate denominator for an F-test of an effect is found as a mean square that under the null hypothesis has the same expectation as the mean square for the effect to be tested. It is readily seen that the main effect for A needs to be tested against Error(A), because for both the mean square has expectaion $\sigma^2 + b\sigma_f^2$ under the null hypothesis of no main effect differences ($H_0 : \alpha_1 = \alpha_2 = ... = 0$). Similarly, it is seen that the main effect for B and the interaction need to be bested against Error(B). The F-statistics are

Main effect A: $$F^{(A)} = \frac{SQ_A/(a-1)}{SS_{error(A)}/[(a-1)(r-1)]}$$

Main effect B: $$F^{(B)} = \frac{SQ_B/(b-1)}{SS_{error(B)}/[a(b-1)(r-1)]}$$

Interaction A × B: $$F^{(AB)} = \frac{SQ_{AB}/[(a-1)(b-1)]}{SS_{error(B)}/[a(b-1)(r-1)]}$$

Example 10 (continued): The data of the rice experiment (Gomez and Gomez, 1984; yields in kg/ha, data slightly modified) were as follows (**rice.dat**):

			Block	
		1	2	3
N1	V1	4520	4208	4030
	V2	4034	5044	3840
	V3	3554	2674	3304
	V4	4216	4212	5016
N2	V1	5598	5256	6162
	V2	6682	5948	5316
	V3	4948	6094	5286
	V4	5372	4694	4382
N3	V1	5806	6600	6794
	V2	5738	6307	6732
	V3	5974	5904	6104
	V4	4276	5924	4236
N4	V1	6192	7146	6860
	V2	6869	7072	6744
	V3	5522	5970	6550
	V4	2504	5126	3818
N5	V1	7470	7578	7642
	V2	7862	6324	6666
	V3	7260	6392	6410
	V4	1594	1690	2856
N6	V1	8542	9012	8548
	V2	6318	7567	5736
	V3	5684	7302	5210
	V4	2338	1560	1744

The ANOVA table is as follows:

Source	d.f.	SS	MS		F	p-value
Main plot stratum:						
Blocks	2	1084820	542410			
N	5	30480453	6096091		10.98	<0.0001
Error (A)	10	5549527	554953	$= E_a$		
Sub plot stratum:						
Variety	3	89885035	29961678		85.74	<0.0001
Interaction	15	69378044	4625203		13.24	<0.0001
Error (B)	36	12579905	349442	$= E_b$		

The interaction is significant, so a comparison of simple N × variety means is useful.

For mean comparisons we also need to account for the two error terms.

Comparison	Standard error of a difference (s.e.d.)	Estimated s.e.d. ($s_{\bar{d}}$)	Degrees of freedom
Marginal means A	$\sqrt{2\left(\frac{\sigma_f^2}{r}+\frac{\sigma^2}{rb}\right)}$	$\sqrt{\frac{2E_a}{rb}}$	$(a-1)(r-1)$
Marginal means B	$\sqrt{\frac{2\sigma^2}{rb}}$	$\sqrt{\frac{2E_b}{ra}}$	$a(b-1)(r-1)$
AB means for constant A	$\sqrt{\frac{2\sigma^2}{r}}$	$\sqrt{\frac{2E_b}{r}}$	$a(b-1)(r-1)$
AB means for constant B	$\sqrt{\frac{2(\sigma_f^2+\sigma^2)}{r}}$	$\sqrt{\frac{2[(b-1)E_b+E_a]}{rb}}$	$df_{Satterthwaite}$

The estimated s.e.d. corresponds to the estimate that is obtained when variance components are estimated by the **ANOVA method** for variance component estimation. In the present case the ANOVA method proceeds by equating the two error mean squares to their expectations and solving for the variance components.

MS	E(MS)
E_a	$\sigma^2 + b\sigma_f^2$
E_b	σ^2

It follows immediately that the sub-plot error variance is estimated as

$\hat{\sigma}^2 = E_b$.

Moreover, subtraction of both mean squares yields

$$E_a - E_b = b\hat{\sigma}_f^2 \Leftrightarrow \hat{\sigma}_f^2 = \frac{E_a - E_b}{b}$$

Plugging these two estimators into the equation for the theoretical s.e.d.'s in the above table (2nd column) yields the estimated s.e.d.'s (third column of table). For example, we find for the AB comparisons for constant B

$$s_{\bar{d}} = \sqrt{\frac{2\left(\hat{\sigma}_f^2 + \hat{\sigma}^2\right)}{r}} = \sqrt{\frac{2\left(b^{-1}\left(E_a - E_b\right) + E_b\right)}{r}} = \sqrt{\frac{2\left(E_a + (b-1)E_b\right)}{rb}}.$$

The s.e.d. can be used to compute a least significant difference as

$$LSD = t_{tab} s_{\bar{d}},$$

where t_{tab} = critical t at chosen significance level.

The degrees of freedom associated with an estimated s.e.d. depend on the error mean square involved. The estimated s.e.d. for comparison of AB means at constant levels of the sub plot factor B then estimated s.e.d. involves a linear combination of ANOVA mean squares. The d.f. in this case can be computed by the so-called Satterthwaite method.

Satterthwaite-method: A single mean square (MS) is distributed as a scaled χ^2-random variable with degrees of freedom depending on the MS. A linear combination of MS has a complicated distribution, but it can be approximated by a χ^2-distribution with the same mean and variance as the linear combination in question (Satterthwaite 1946). This approximating χ^2-distribution has degrees of freedom that depend on the degrees of freedom of the MS involved in the linear combination. The degrees of freedom of the linear combination of MS

$$Q = c_1 MS_1 + c_2 MS_2 + ...$$

is computed as

$$df_{Satterthwaite} = \frac{\left(c_1 MS_1 + c_2 MS_2 + ...\right)^2}{\frac{\left(c_1 MS_1\right)^2}{df_1} + \frac{\left(c_2 MS_2\right)^2}{df_2} + ...},$$

where df_i is the degrees of fredom of MS_i.

In case of a split plot design with main plots laid out in complete blocks the approximate d.f. for the standard error $s_{\bar{d}} = \sqrt{\frac{2\left[(b-1)E_b + E_a\right]}{rb}}$ are:

$$df_{Satterthwaite} = \frac{\left(\frac{2E_a}{rb} + \frac{2(b-1)E_b}{rb}\right)^2}{\frac{\left(\frac{2E_a}{rb}\right)^2}{(a-1)(r-1)} + \frac{\left(\frac{2(b-1)E_b}{rb}\right)^2}{a(b-1)(r-1)}}$$

Example 11 (continued): For the rice experiment we find

$E_a = 554953,\ E_b = 349442$
$a = 6,\ b = 4,\ r = 3$

and thus

$$s_{\bar{d}} = \sqrt{\frac{2[(b-1)E_b + E_a]}{rb}} = \sqrt{\frac{2[3 \times 349442 + 554953]}{12}} = 516.93 \quad \text{and}$$

$$df_{Satterthwaite} = \frac{\left(\frac{2E_a}{rb} + \frac{2(b-1)E_b}{rb}\right)^2}{\frac{\left(\frac{2E_a}{rb}\right)^2}{(a-1)(r-1)} + \frac{\left(\frac{2(b-1)E_b}{rb}\right)^2}{a(b-1)(r-1)}} = \frac{\left(\frac{2 \times 554953}{12} + \frac{2 \times 3 \times 349442}{12}\right)^2}{\frac{\left(\frac{2 \times 554953}{12}\right)^2}{10} + \frac{\left(\frac{2 \times 3 \times 349442}{12}\right)^2}{36}} = 41.9$$

Using a mixed model package we find the following results:

Covariance Parameter Estimates

Cov Parm	Estimate
block*n	51378
Residual	349442

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	2	10	0.98	0.4095
n	5	10	10.98	0.0008
v	3	36	85.74	<.0001
n*v	15	36	13.24	<.0001

The F-tests are Wald-type F-tests and are identical to those in the ANOVA table given previously. Thus, the Wald-type F-tests are the same as those obtained by the usual sequential sum of squares. This coincidence is a result of the balanced data structure and will not hold for unbalanced data. The interactions are significant, so we compare N × variety means. Some examples are given below.

Least Squares Means

Effect	n	v	Estimate	Standard Error
n*v	1	1	4252.67	365.52
n*v	1	2	4306.00	365.52
n*v	1	3	3177.33	365.52
n*v	1	4	4481.33	365.52
n*v	2	1	5672.00	365.52
n*v	2	2	5982.00	365.52
n*v	2	3	5442.67	365.52
n*v	2	4	4816.00	365.52
n*v	3	1	6400.00	365.52
n*v	3	2	6259.00	365.52
n*v	3	3	5994.00	365.52
n*v	3	4	4812.00	365.52
n*v	4	1	6732.67	365.52
n*v	4	2	6895.00	365.52
n*v	4	3	6014.00	365.52
n*v	4	4	3816.00	365.52
n*v	5	1	7563.33	365.52
n*v	5	2	6950.67	365.52
n*v	5	3	6687.33	365.52
n*v	5	4	2046.67	365.52
n*v	6	1	8700.67	365.52
n*v	6	2	6540.33	365.52
n*v	6	3	6065.33	365.52
n*v	6	4	1880.67	365.52

Differences of Least Squares Means

Effect	n	v	_n	_v	Estimate	Standard Error	DF	t Value	Pr > |t|
n*v	1	1	1	2	-53.3333	482.66	36	-0.11	0.9126
n*v	1	1	2	1	-1419.33	516.93	41.9	-2.75	0.0089

<more comparisons>

Type of comparison	s.e.d.	§LSD(α = 5%)
N × variety means at constant N	**482.66**	**978.88**
N × variety means at constant variety	**516.93**	**1043.26**

§ $LSD = t_{tab} s_{\bar{d}}$

The means can be displayed as follows:

	N1 (0)	N2 (60)	N3 (90)	N4 (120)	N5 (150)	N6 (180)
V1	4253	5672	6400	6733	7563	8701
V2	4306	5982	6259	6895	6951	6540
V3	3177	5443	5994	6014	6687	6065
V4	4481	4816	4812	3816	2047	1881

LSD(α = 5%) = 1043.26 (comparisons within a row)
LSD(α = 5%) = 978.8 (comparisons within a column)

Unbalanced data: The ANOVAs based on sequential sums of squares we have shown produce F-tests that are identical to those produced by the Wald-type F-tests for mixed models, when data are balanced and REML estimates of variance components are positive. When data are unbalanced, an ANOVA table can still be produced, but construction of F-tests is more difficult, because the expected mean squares do not have a simple form and linear combinations of several mean squares are needed to construct appropriate error terms. Alternatively, a REML analysis will always produce a Wald-type F-test that is approximately valid. It is one of the great advantages of REML analysis of a mixed model, that unbalanced data pose no problem. With unbalanced data, standard errors of a difference (s.e.d.) will not be constant among pairs of treatments, but one can report am ean s.e.d. for descriptive purposes.

The equations given for s.e.d. and associated the Satterthwaite degrees of freedom assume balanced data as well. For balanced data, treatment means are simple means and squared standard errors can be expressed as linear combinations of mean squares, which is the basis for the Satterthwaite method. For unbalanced data, these simple settings no longer apply. For example, adjusted means must be used. The more general Kenward-Roger (1997) method may be applied in this situation to both estimate standard errors and compute approximate degrees of freedom.

SAS hints

```
data rice;
input
block    n    v    n_amount    yield;
datalines;
      1      1    1         0       4520
      1      1    2         0       4034
<more data>
      3      6    2       180       5736
      3      6    3       180       5210
      3      6    4       180       1744
;
proc mixed data=rice;
class block n v;
model yield=block n v n*v / ddfm=KR;
random n*block; /*main plot error*/
lsmeans n*v / pdiff;
run;

proc glimmix data=rice;
class block n v;
model yield=block n v n*v / ddfm=KR;
```

```
random n*block; /*main plot error*/
lsmeans n*v / pdiff;
slice n*v / sliceby=n lines; /*because we have interaction*/
slice n*v / sliceby=v lines; /*because we have interaction*/
run;
```

5.3 Polynomial regression for a split-plot experiment

When at least one of the factors in a factorial experiment is quantitative, one should consider polynomial regression, as discussed for single factor experiments. For two quantitative factors, it is useful to look at so-called **response surface methodology** (see book by Dean and Voss, 1999; also see Mead et al., 1993, Section 16.7).

Example 11 (continued): The N-levels for the six N-treatments in the rice data (**rice.dat**) were as follows:

N	N-dose	Variate in model
1	0	x_1
2	60	x_2
3	90	x_3
4	120	x_4
5	150	x_5
6	180	x_6

We can consider fitting a polynomial. The ANOVA has revealed significant interaction between variety and N-level, so the regression curves are not parallel. Thus, we consider fitting the following model:

$$y_{ijh} = \mu + b_h + \beta_j + \gamma_{j1}x_i + f_{ih} + e_{ijh} \quad ,$$

where

γ_{j1} = slope of j-th variety for the linear regression on N-dose
x_i = i-th N-dose

This model implies that each variety has its own regression line. To test the lack-of-fit, we need a deviation from the regression for each variety at each N-level. Thus, the lack-of-fit term needs to be indexed by both i and j. The model is

$$y_{ijh} = \mu + b_h + \beta_j + \gamma_{j1}x_i + \delta_{ij} + f_{ih} + e_{ijh}$$

where

δ_{ij} = lack-of-fit effect

We may add polynomial terms until the lack-of-fit test is not significant. For example, the lack-of-fit test for the quadratic model is obtained by the following model:

$$y_{ijh} = \mu + b_h + \beta_j + \gamma_{j1}x_i + \gamma_{j2}x_i^2 + \delta_{ij} + f_{ih} + e_{ijh}$$

To fit these models, we need to generate a variable for the lack-of-fit effect (δ_{ij}), which has the same levels as the factor for N-fertilizer.

block	n	v	n_amount	lackfit	yield
1	1	1	0	1	4520
1	1	2	0	1	4034
1	1	3	0	1	3554
1	1	4	0	1	4216
1	2	1	60	2	5598
.					
.	< more data >				
.					
3	5	3	150	5	6410
3	5	4	150	5	2856
3	6	1	180	6	8548
3	6	2	180	6	5736
3	6	3	180	6	5210
3	6	4	180	6	1744

BLOCK, N and V are the variables for blocks, N-level and variety. N is the actual amount of fertilizer applied and is the quantitative variable x_j for the regression. To fit the quadratic term, we may further generate a variable N2 defined as the squared N-level (N2). The variable LACKFIT is a qualitative factor that has the same levels as N and is needed for fitting the lack-of-fit effects. The key model terms may then be represented as

$\gamma_{j1}x_i$: **V*N_amount**
$\gamma_{j2}x_i^2$: **V*N_amount*N_amount**
δ_{ij} : **V*Lackfit**

We first fit the linear term and test for lack-of-fit.

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	2	10	0.98	0.4095
v	3	36	85.74	<.0001
n_amount*v	4	36	52.59	<.0001
lackfit*v	16	36	2.69	0.0068

The lack-of-fit is significant, so we add a quadratic term:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	2	10	0.98	0.4095
v	3	36	85.74	<.0001
n_amount*v	4	36	52.59	<.0001
n_amount*n_amount*v	4	36	8.11	<.0001
lackfit*v	12	36	0.89	0.5665

Now the lack of fit is not significant, i.e., the quadratic response fits well. The remaining task is to fit the quadratic regression curves for each variety. To do this, we fit the quadratic model without the lack-of-fit effect:

$$y_{ijh} = \mu + b_h + \beta_j + \gamma_{j1}x_i + \gamma_{j2}x_i^2 + f_{ik} + e_{ijh}$$

The expected value of an observation is obtained from the model by "stripping off" the random effects:

$$\eta_{ijh} = \mu + b_h + \beta_j + \gamma_{j1}x_i + \gamma_{j2}x_i^2$$

Obviously, the intercept depends not only on varieties, but also on blocks. Thus, we may compute the average across blocks for convenience:

$$\overline{\eta}_{ij\bullet} = \mu + \overline{b}_\bullet + \beta_j + \gamma_{j1}x_i + \gamma_{j2}x_i^2$$

The intercept of the j-th variety is

$$\mu + \overline{b}_\bullet + \beta_j$$

From the results, the estimated quadratic regression equations are:

Variety V1: YIELD = 4323 + 18.81 × N + 0.02583 × N^2
Variety V2: YIELD = 4279 + 35.62 × N – 0.12481 × N^2
Variety V3: YIELD = 3201 + 45.29 × N – 0.16049 × N^2
Variety V4: YIELD = 4562 + 14.61 × N – 0.17638 × N^2

The fitted curves are displayed in Fig. 6.1.

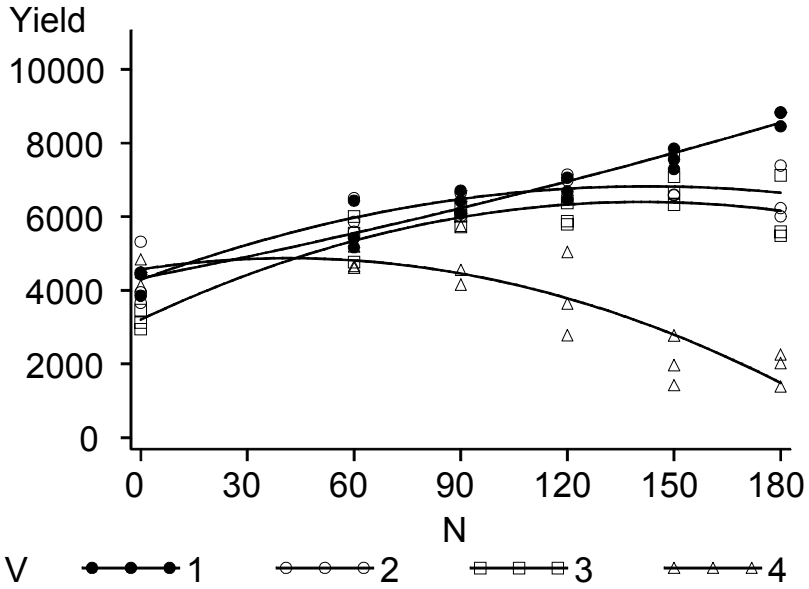


Fig. 4: Fitted regression curves for varieties V1 to V4.

V	Variety
1	IR8
2	IR5
3	C4-63
4	Peta

(Gomez & Gomez, 1984, p.102)

Chrispeels MJ & DE Sadava. (file of photo kindly provided by Ignacio Romagosa) Plants, Genes and Agriculture Plants, Genes and Crop Biotechnology Jones and Bartlett Publ.

The long-strawed variety Peta (V4) is an old variety that is susceptible to lodging, while the other varieties such as IR8 (V1) are short-strawed and less susceptible to lodging. This is nicely reflected in the fitted regression curves.

Are curves parallel?

To test if the curves of the variaties run parallel, we need to consider the linear and quadratic regression terms. If the curves are parallel, they can be modelled by

$$\eta_{ijh} = \mu + b_h + \beta_j + \phi_1 x_i + \phi_2 x_i^2 \ ,$$

where ϕ_1 is a common linear term and ϕ_2 is a common quadratic regression term. To compare this to the model where each variety has its own linear regression term γ_{j1} and its own quadratic regression term γ_{j2}, we can express the variety-specific model in such a way that

the model with common linear and quadratic terms is contained as a special case. Thus, we use the re-parametrizations

$\gamma_{j1} = \phi_1 + \alpha_{j1}$ and
$\gamma_{j2} = \phi_2 + \alpha_{j2}$,

where α_{j1} and α_{j2} are linear and quadratic interaction terms. The null hypothesis that the curves are parallel can now be translated into the null hypothesis

$H_0 : \alpha_{1j} = \alpha_{2j} = 0$ for all varieties *j*. Plugging the re-parameterization into our model, we obtain the model

$\eta_{ijh} = \mu + b_h + \beta_j + \phi_1 x_i + \alpha_{2j} x_i + \phi_2 x_i^2 + \alpha_{1j} x_i^2$.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	2	13	1.16	0.3425
v	3	45	3.30	0.0286
n_amount	1	13	40.20	<.0001
n_amount*v	3	45	3.33	0.0278
n_amount*n_amount	1	13	20.82	0.0005
n_amount*n_amount*v	3	45	4.91	0.0049

The F-tests show that there is significant heterogeneity between varieties for both the linear term (p = 0.0278) and the quadratic term (p = 0.0049). In the above analysis, the terms were coded as follows:

Model term	Coding in SAS
β_j	v
$\phi_1 x_i$	n_amount
$\phi_2 x_i^2$	n_amount*n_amount
$\alpha_{1j} x_i$	v*n_amount
$\alpha_{2j} x_i^2$	v*n_amount*n_amount

SAS hints

Example code shows how to do lack-of-fit test for linear model and how to report the quadratic model, for which the lack-of-fit is non-significant (code not shown).

```
data rice;
set rice;
lackfit=n;
run;
```

```
/*lack-of-fit for linear model*/
proc print data=rice;
run;

proc mixed data=rice;
class block n v lackfit;
model yield=block v v*n_amount v*lackfit
             / ddfm=KR solution htype=1;
random n*block; /*main plot error*/
run;

/*Are curves parallel?*/
proc mixed data=rice;
class block n v;
model yield=block v n_amount          v*n_amount
                    n_amount*n_amount v*n_amount*n_amount/ ddfm=KR;
random n*block; /*main plot error*/
run;

/*report quadratic model*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
estimate 'intercept v=1' intercept 3 block 1 1 1 v 3 0 0 0 / divisor=3;
estimate 'intercept v=2' intercept 3 block 1 1 1 v 0 3 0 0 / divisor=3;
estimate 'intercept v=3' intercept 3 block 1 1 1 v 0 0 3 0 / divisor=3;
estimate 'intercept v=4' intercept 3 block 1 1 1 v 0 0 0 3 / divisor=3;
run;
```

6 Repeated measures

6.1 The duckweed experiment

Example 12: *Lemna* species, also known as duckweed, grow as simple free-floating thalli on or just beneath the water surface. The plants grow mainly by vegetative reproduction: two daughter plants bud off from the adult plant. This form of growth allows very rapid colonisation of new water (wikipdia.org).

An experiment was performed to compare the growth of five different groups of *Lemna* species (Thöni, 1970; **lemna.dat**). The groups differed in the concentration of Glycolstearate in the nutrient solution. This chemical was added to reduce water activity and thus the water uptake by the plants. Each group consisted of six plants. For each plant, the number of daughter plants per individual was recorded at three time points (1, 5 and 7 weeks). The counts were $\log_{10}$ transformed and multiplied by 10. A plot of the transformed data against time is shown in Fig. 5. It appears from the plot that there are some differences between groups in terms of the profiles (average level and growth over time). An objective of the analysis was to find out whether there were significant differences between groups in the profiles or growth curves.

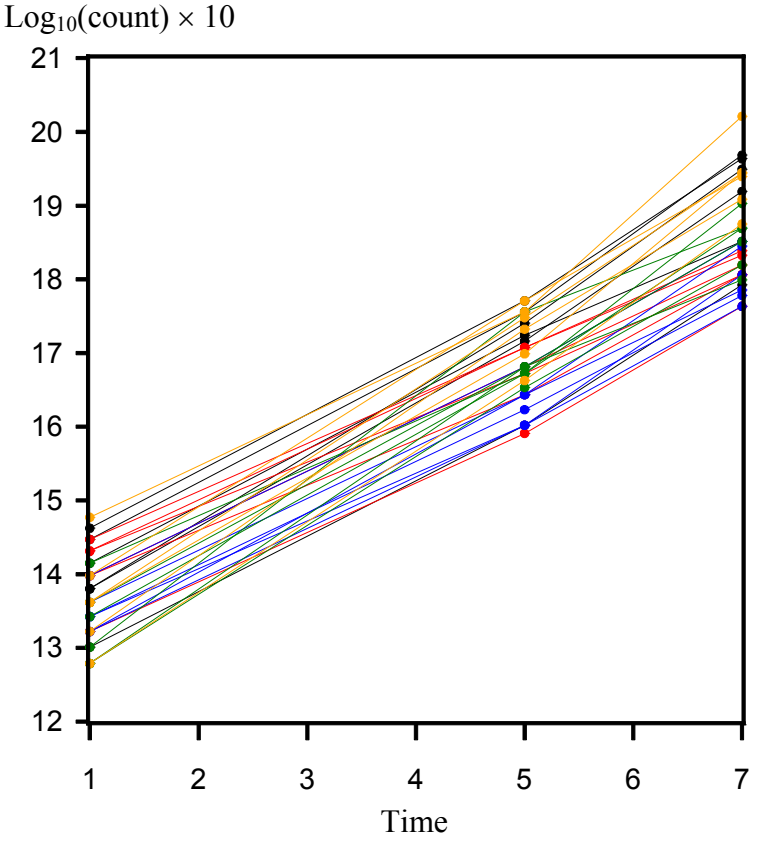


Fig. 5: Plot of $\log_{10}$(count) (× 10) versus time for 30 individuals and three repeated measurements per individual. Five different groups are marked by different colors.

Thus the factors of interest are group and time. We could run a standard two-way ANOVA based on the model

$$E(y_{ijh}) = \eta_{ijh} = \mu + \tau_{ij}$$

where

y_{ijh} = observation of h-th individual in i-th group an j-th time point
μ = general effect (intercept)
τ_{ij} = effect of i-th group and j-th time point

The treatment effect τ_{ij} can be factorized as usual:

$$\tau_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

The problem is that we cannot just use the model

$$y_{ijh} = \mathrm{E}(y_{ijh}) + e_{ijh}$$

with independent errors $e_{ijh} \sim \mathrm{N}(0, \sigma^2)$, because we have repeated measurements taken in the same individual. The repeated measurements will cause correlations among observations taken from the same individual (see Fig. 6).

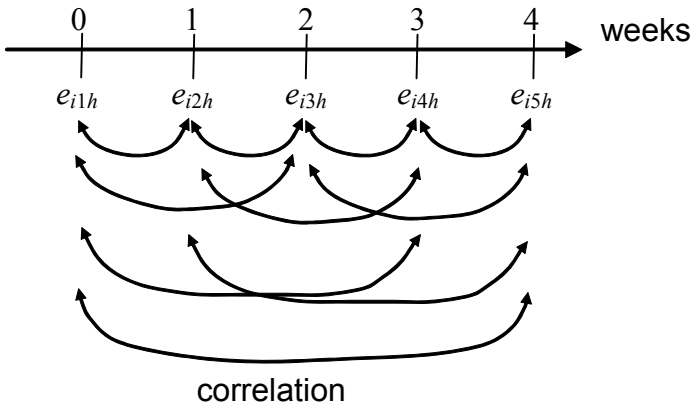


Fig. 6: Schematic representation of correlation among errors of repeated measurements on the same experimental unit.

The situation is in some ways comparably to a split-plot design, if we regard individuals as main plots and repeated measurements as sub plots. In a split-plot experiment, observations on the same main plot are positively correlated due to the common main plot error term for all subplots. As a result of the simple variance structure, the correlation is the same for all pairs of sub plots. By contrast, with repeated measurements it is seldom suitable to assume that all pairs of time points are equally correlated. Typically, correlation decays with distance in time.

The main difference between a split plot design and repeated measurements lies in the fact that time points cannot be randomized, whereas sub plots in a split-plot experiment can be randomized. Randomization theory justifies the assumption of independent sub plot errors within a main plot, and of equal correlation of sub plots within the same main plot. By contrast, lack of randomization for repeated measurements usually renders the equal correlation model unsuitable.

There are many models for correlation on longitudinal data (repeated measurements over time). One popular model is the **autoregressive AR(1) model**, under which the correlation among the observations at time points j and j' on the ih-th unit is (dropping other indices for simplicity)

$$\mathrm{cov}\left(e_j, e_{j'}\right) = \sigma^2 \rho^{|j-j'|} \quad (0 < \rho < 1),$$

where ρ is the autocorrelation parameter. It is assumed here that the index j is in time order, so $j = 1$ is the first time point, $j = 2$ is the second, etc. Thus

$$\mathrm{cov}\left(e_1, e_1\right) = \sigma^2 \rho^{|1-1|} = \sigma^2 \qquad (\rho^0 = 1!)$$
$$\mathrm{cov}\left(e_1, e_2\right) = \sigma^2 \rho^{|1-2|} = \sigma^2 \rho$$
$$\mathrm{cov}\left(e_1, e_3\right) = \sigma^2 \rho^{|1-3|} = \sigma^2 \rho^2$$
$$\mathrm{cov}\left(e_1, e_4\right) = \sigma^2 \rho^{|1-4|} = \sigma^2 \rho^3$$

etc.

The larger the distance in time, the lower is the covariance and correlation. This is shown in Fig. 7 for a few values of the autocorrelation ρ.

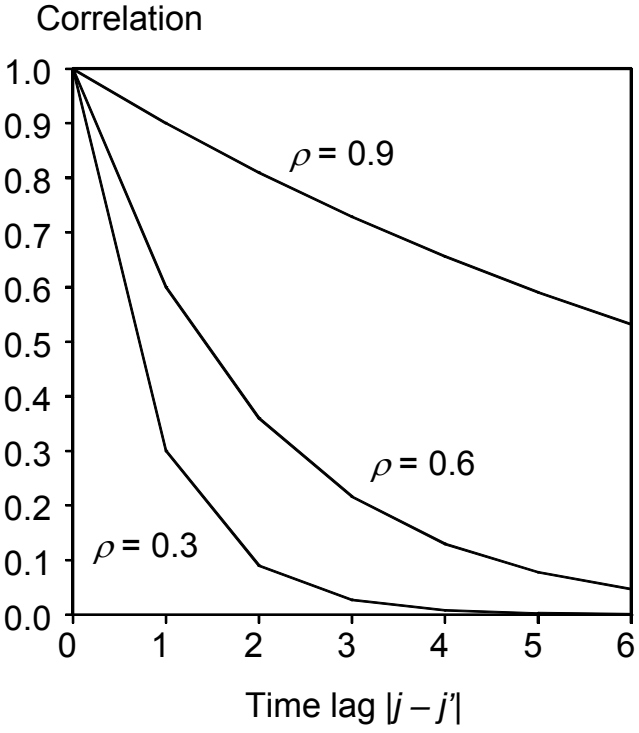


Fig. 7: Correlation under an AR(1) model as a function of time lag for different values of the autocorrelation ρ.

This correlation model is useful, if all time points are equally spaced. With unequal spacing, the model could be expended as follows:

$$\mathrm{cov}\left(e_j, e_{j'}\right) = \sigma^2 \rho^{d(j,j')}$$

where $d(j, j')$ = time lag between time points j and j'.

There are many alternative models, which allow the corelation to decay with distance in time (Schabenberger and Pierce, 2002). The most general model is to let both the variance at a time point and the covariances at a time point vary freely. This **unstructured (UN)** model can be

represented as

$$\text{cov}(e_j, e_{j'}) = \sigma_j \sigma_{j'} \rho_{jj'} \ ,$$

where

σ_j^2 = variance at time j and
$\rho_{jj'}$ = correlation between time points j and j'

At the other extreme, we could postulate equal variance and correlation, which is known as the **compound symmetry (CS) model** or **equal correlation model**, such that

$$\text{cov}(e_j, e_{j'}) = \sigma^2 \rho \quad (j \neq j) \quad \text{and}$$

$$\text{var}(e_j) = \sigma^2 \ .$$

The CS model is also known as **split-plot in time model**, because it is equivalent to a split-plot model with individuals as main plots and time points within individuals as sub plots.

And, of course, one may also consider the **independent model (ID)** with homogeneous variances, though this is not usually realistic.

$$\text{cov}(e_j, e_{j'}) = 0 \quad (j \neq j) \quad \text{and}$$

$$\text{var}(e_j) = \sigma^2 \ .$$

Now there are several alternative models for the variance-covariance structure, which raises a model selection problem. Model selection may be based on likelihood-based methods.

6.2 Model selection by likelihood ratio tests and AIC

The basis for both the likelihood ratio test and the Akaike Information Criterion (AIC) is the log-likelihood of the fitted model.

Example 12 (continued): Fitting the models described above models to the duckweed data (**lemna.dat**) we find the following results:

Model for e_{ijh}	$-2 \times$ restricted log-likelihood
Independent (ID)	128.3
Compound symmetry (CS)	84.9
Autoregressive AR(1)	96.9
Unstructured (UN)	79.3

The table gives minus twice the restricted log-likelihood ($-2 \log L_R$), the so-called **deviance**. This will be convenient for the likelihood ratio test. It is seen that the unstructured model has the smallest deviance ($-2 \log L_R$), and so the largest restricted log-likelihood. This alone does

not mean that unstructured is the best model. It can be shown that the log-likelihood must increase as we add parameters to the model. This is analogous to the reduction in error sum of squares when a parameter is added to the linear model. To see whether the unstructured model is really the best model, it must be compared to the other models by likelihood-based methods.

Before coming to such methods, first consider the fit of the unstructured model to the duckweed data. We find the following parameter estimates:

Parameter	Estimate
σ_1^2	0.2739
σ_2^2	0.1885
σ_3^2	0.2170
ρ_{12}	0.6552
ρ_{13}	0.8268
ρ_{23}	0.7286

There does not seem to be much heterogeneity in the variances of the three time points. Also, the correlations are relatively homogeneous. Strikingly, the most distant time points have the largest correlation. This is in contrast to models, where correlation decays with lag distance, which makes these models somewhat implausible for the data at hand. We will verify this by likelihood-based methods of inference.

Likelihood ratio tests can be applied only to **nested models**. Two models are nested if one the one model, known as the **reduced model,** can be regarded as a special case of the other model, known as the **full model**. For example, in a one-way analysis of variance we consider the full model

$y_{ij} = \mu + \alpha_i + e_{ij}$.

Under the null hypothesis of equal treatments, i.e., $H_0 : \alpha_1 = \alpha_2 = ... = 0$ we obtain the reduced model

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

which states that all treatments have the same mean μ. Under the alternative hypothesis H_A, each treatment has its own mean ($\mu + \alpha_i$).

Likelihood ratio test: A general principle says that a null hypothesis H_0 can be tested against the alternative hypothesis H_A by comparing the likelihoods of the corresponding full and the reduced model, i.e., by the likelihood ratio

$$LR = \frac{L(H_0)}{L(H_A)} ,$$

where $L(H)$ is the likelihood of the model under hypothesis H. It is more convenient to work with the logarithm

$$\log LR = \log L(H_0) - \log L(H_A) \ .$$

Under the null hypothesis H_0 the statistic

$$T = -2 \log LR$$

asymptotically for large samples has a chi-squared distribution with $v = p_{HA} - p_{H0}$ degrees of freedom, where

p_{HA} = the number of parameters under the alternative hypothesis H_A and
p_{H0} = the number of parameters under the null hypothesis H_0.

Interestingly, applying this test to the one-way ANOVA situation yields a test that is equivalent to the usual ANOVA F-test! But the test is more generally applicable, which is the main reason for its popularity.

Example 12 (continued): We now use the test to compare the various variance-covariance models to the unstructured model. Note that all models are special cases of the unstructured model. Thus, all models are reduced models compared to the unstructured full model. The models are nested in the unstructured model and so can be compared by an LR test.

Model for e_{ijh}	No. of parameters p	$-2 \times \log L_R$	T relative to unstructured	d.f. for T v
Independent (ID)	1	128.3	$^{sign}49.0$	5
Compound symmetry (CS)	2	84.9	$^{ns}5.6$	4
Autoregressive AR(1)	2	96.9	$^{sign}17.6$	4
Unstructured (UN)	6	79.3	-	-

The 5% critical values for the chi-squared distribution are 9.49 for $v = 4$ d.f. and 11.07 for $v = 5$ d.f. Thus, the compound symmetry model is not significantly different from the unstructured model, but the independent and AR(1) models are different. It may be concluded that the compound symmetry model is the best model. Its advantage compared to the unstructured model is the smaller number of parameters.

A limitation of the LR test is that we can only compare nested model. In the duckweed example, the independent model is nested within the AR(1) and CS models, but we cannot compare the AR(1) and CS models by an LR test, because these models are not nested.

A simple method that can also be applied to non-nested models is to compute model selection criteria such as the **Akaike Information Criterion (AIC)**. This is computed as

$$AIC = -2 \log L_R + 2p \ ,$$

where L_R is the restricted log likelihood and p is the number of variance parameters of the model. **The smaller AIC the better is the model**.

Example 12 (continued): For the duckweed example, the CS model has the smallest AIC and so is preferred.

Model for e_{ijh}	p	$-2 \times \log L_R$	AIC
Independent (ID)	1	128.3	130.3
Compound symmetry (CS)	2	84.9	88.9
Autoregressive AR(1)	2	96.9	100.9
Unstructured (UN)	6	79.3	91.3

Caution: In order to compare different variance-covariance structures using the restricted log-likelihood, one must have **the same fixed effects structure**.

SAS hints

```
data lemna;
input
grp    ind    time       y       count;
plant=ind;
datalines;
                          1      11      1     14.1497      26
                          1      11      5     17.2428      53
                          1      11      7     18.5126      71
<more data>
                          5      56      1     13.9794      25
                          5      56      5     17.7085      59
                          5      56      7     19.3952      87
;
/*AR(1)*/
proc mixed data=lemna;
class grp time plant;
model y=grp time grp*time /ddfm=KR;
repeated time/subject=plant type=ar(1);
run;

/*CS*/
proc mixed data=lemna;
class grp time plant;
model y=grp time grp*time /ddfm=KR;
repeated time/subject=plant type=CS;
run;

/*UN = in terms of covariances*/
proc mixed data=lemna;
class grp time plant;
model y=grp time grp*time /ddfm=KR;
repeated time/subject=plant type=UN;
run;

/*UNR = in terms of correlations*/
proc mixed data=lemna;
class grp time plant;
model y=grp time grp*time /ddfm=KR;
repeated time/subject=plant type=UNR;
run;

/*VC = independent model*/
proc mixed data=lemna;
```

```
class grp time plant;
model y=grp time grp*time /ddfm=KR;
repeated time/subject=plant type=VC;
run;

/*plot of data*/

symbol i=join value=dot;
proc gplot data=lemna;
plot y*time=plant;
run;
```

6.3 Inference for fixed effects using selected variance-covariance model

Using the selected variance-covariance model, we can now perform Wald-tests for the fixed effects. For the duckweed data we find:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	2	50	3098.74	<.0001
grp	4	25	4.10	0.0109
grp*time	8	50	10.87	<.0001

There is evidence of group × time interaction, so the profiles are not the same between groups.

To further explore the profiles, we may consider fitting a straight line for each group. The regression model is

$$\tau_{ij} = \alpha_i + \gamma_i t_j ,$$

where t_j is the time at time point j. As there are replicate data, we can test the lack-of-fit. To test the lack of fit, we add a lack-of-fit term δ_{ij}:

$$\tau_{ij} = \alpha_i + \gamma_i t_j + \delta_{ij}$$

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
grp	4	25	4.10	0.0109
time*grp	5	50	1255.22	<.0001
grp*lackfit	5	50	1.67	0.1598

There is no significant lack of fit, so the linear regression seems appropriate. To test for differences in regression slopes, we drop the lack-of-fit term and add a general regression term:

$$\tau_{ij} = \alpha_i + \beta t_j + \gamma_i t_j$$

The term $\gamma_i t_j$ now is an interaction term due to the presence of the general regression term βt_j, and hence the interaction models heterogeneity in slopes.

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
grp	4	25	4.10	0.0109
time	1	55	5838.83	<.0001
time*grp	4	55	19.60	<.0001

There is a significant heterogeneity in slopes. The fitted model is depicted in Fig. 8.

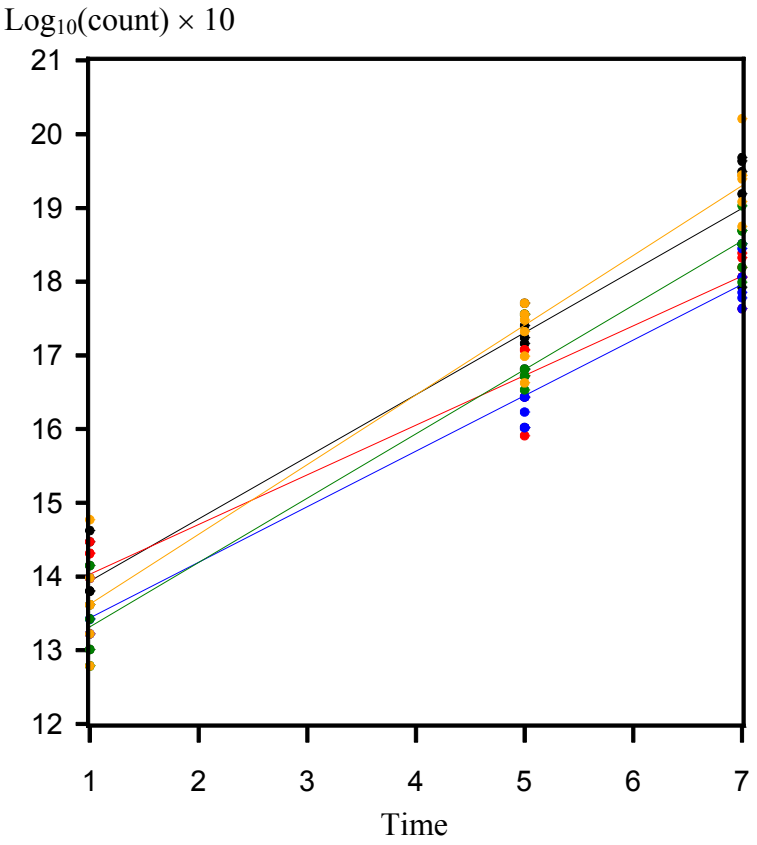


Fig. 8: Plot of fitted linear regression models based on mixed model with CS structure for residual error of repeated measurements. Groups indicated by different colours.

The fitted parameters of the linear regression as per the mixed model with a compound symmetry structure for residual errors are as follows:

Solution for Fixed Effects

Effect	grp	Estimate	Standard Error	DF	t Value	Pr > |t|
grp	1	13.0927	0.2044	43	64.05	<.0001
grp	2	13.3541	0.2044	43	65.32	<.0001
grp	3	12.6845	0.2044	43	62.05	<.0001
grp	4	12.4400	0.2044	43	60.85	<.0001
grp	5	12.6764	0.2044	43	62.01	<.0001
time*grp	1	0.8427	0.02395	55	35.19	<.0001
time*grp	2	0.6744	0.02395	55	28.16	<.0001
time*grp	3	0.7541	0.02395	55	31.49	<.0001
time*grp	4	0.8735	0.02395	55	36.48	<.0001
time*grp	5	0.9467	0.02395	55	39.54	<.0001

Remark: More complex models than just the linear can, of course, be used to model profiles. This would usually require more than three time points. If an intrinsically non-linear model is used, a non-linear mixed model results, for which speciall fitting algorithms are needed. For details see, e.g. Schabenberger and Pierce (2002).

SAS hints

```
/*regression analysis*/

data lemna;
set lemna;
time_class=time;
run;

proc mixed data=lemna;
class grp plant time_class;
model y=grp grp*time /ddfm=KR solution noint;
                                          /*noint supresses overall mu*/
repeated time_class/subject=plant type=CS;
run;

/*lack of fit*/

data lemna;
set lemna;
lackfit=time;
run;

  /*option 1*/
proc mixed data=lemna;
class grp plant time_class;
model y=grp grp*time grp*time_class/ddfm=KR;
repeated time_class/subject=plant type=CS;
run;

  /*option 2*/
proc mixed data=lemna;
class grp plant time_class lackfit;
model y=grp grp*time grp*lackfit/ddfm=KR;
repeated time_class/subject=plant type=CS;
```

```
run;

/*test for equality of slopes*/
proc mixed data=lemna;
class grp plant time_class;
model y=grp time grp*time /ddfm=KR solution noint;
repeated time_class/subject=plant type=CS;
run;
```

6.4 A more complex example with *Arabidopsis*

Example 13: A genetic experiment was performed with NILs derived from an Arabidopsis cross (**arabidopsis.dat**). Each NIL was tested for *per se* performance. In addition, a NIL was backcrossed to each of its parents and to the F1 of the original cross. These crosses were made to fit a specific quantitative-genetic model, which will not be considered here (Melchinger et al., 2007). For each NIL, there were four genotypes, i.e., the NIL itself and its three crosses. The experiment was laid out as a split-plot design with NILs as main plot factor and genotypes within NILs as sub plot factor. Main plots were laid out according to an α-design.

At three different time points, 15, 22 and 29 days after emergence (DAE), the rosette diameter was recorded. Thus, we have three repeated measurements per subplot. Before developing a model for repeated measures, we consider the model for analysis of a single point in time. To represent the block structure, we need effects for replicate, block within replicate, and main plot within blocks. Errors will correspond to sub plots within main plots. The treatment model will just comprise a simple main effect for genotypes. There is more structure in the genotypes, but this will be ignored for the moment. The model is

$$y_{ijkhp} = \mu + \alpha_i + r_j + b_{jk} + m_{jkh} + s_{jkhp},$$

where y_{ijkhp} is the measurement for the i-th treatment (genotype) in the j-th replicate, k-th block within j-th replicate, h-th main plot within jk-th block and p-th subplot within jkh-th main plot. The effects represent:

α_i: i-th treatment (genotype)
r_j: j-th replicate
b_{jk}: jk-th block (random)
m_{jkh}: jkh-th main plot (random)
s_{jkhp}: $jkhp$-th sub plot (random)

Now we need to extend the model to repeated measures. To derive a model, it is reasonable to postulate that the model is commensurate with the one we postulate for a single time point. This can be achieved by just adding an index t for time points.

$$y_{ijkhpt} = \mu_t + \alpha_{it} + r_{jt} + b_{jkt} + m_{jkht} + s_{jkhpt}.$$

To model the repeated measures variance-covariance structure, it is convenient to collect correlated random effects for randomization units (blocks, main plots and sub plots) into vectors. The vectors are

$$\boldsymbol{b}_{jk} = \begin{pmatrix} b_{jk1} \\ b_{jk2} \\ . \\ . \\ b_{jkT} \end{pmatrix}, \boldsymbol{m}_{jkh} = \begin{pmatrix} m_{jkh1} \\ m_{jkh2} \\ . \\ . \\ m_{jkhT} \end{pmatrix}, \text{and } \boldsymbol{s}_{jkhp} = \begin{pmatrix} s_{jkhp1} \\ s_{jkhp2} \\ . \\ . \\ s_{jkhpT} \end{pmatrix} .$$

For these, we may postulate repeated-measures $T \times T$ variance-covariance matrices:

$$\text{var}(\boldsymbol{b}_{jk}) = \boldsymbol{\Sigma}_b$$
$$\text{var}(\boldsymbol{m}_{jkh}) = \boldsymbol{\Sigma}_m$$
$$\text{var}(\boldsymbol{s}_{jkhp}) = \boldsymbol{\Sigma}_s$$

where the structures may be chosen as AR(1), CS, unstructured etc. All vectors are uncorrelated among themselves. For example, $\boldsymbol{b}_{11}$ is uncorrelated with $\boldsymbol{b}_{12}$ and $\boldsymbol{b}_{23}$. More formally

$$\text{cov}(\boldsymbol{b}_{jk}, \boldsymbol{b}_{j'k'}) = \boldsymbol{0} \qquad \text{for } j \neq j \text{ or } k \neq k'$$
$$\text{cov}(\boldsymbol{m}_{jkh}, \boldsymbol{m}_{j'k'h'}) = \boldsymbol{0} \qquad \text{for } j \neq j \text{ or } k \neq k' \text{ or } h \neq h'$$
$$\text{cov}(\boldsymbol{s}_{jkhp}, \boldsymbol{s}_{j'k'h'p'}) = \boldsymbol{0} \qquad \text{for } j \neq j \text{ or } k \neq k' \text{ or } h \neq h' \text{ or } p \neq p'$$

A plot of rosette diameter versus time, with the time variable randomly displaced along the abscissa on an interval of ±0.2 around the true time point for better visualisation, shows that heterogeneity of variance among time points is very dramatic (Fig. 9). This suggests that a covariance structure that allows for heterogeneity of variance, is preferable. Thus, we here also consider a modification of the independent (ID), CS and AR(1) models that cater for heterogeneity of variance, i.e. the DIAG, CSH and the ARH(1) structures, respectively. If $\boldsymbol{C}$ denotes the structure for ID, CS and AR(1), then the extension is $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}$ for DIAG, CSH and ARH(1), respectively, where $\boldsymbol{D}$ is a diagonal matrix with time-specific variances on the diagonal:

$$\boldsymbol{D} = \begin{pmatrix} \sigma_1^2 & & & & \\ & \sigma_2^2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_T^2 \end{pmatrix}, \boldsymbol{D}^{1/2} = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_T \end{pmatrix}.$$

In the following, the correlation matrices for various models are given.

ID:

$$\boldsymbol{C}=\begin{pmatrix}1 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & & \vdots\\ 0 & 0 & \ddots & \ddots & 0\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & 0 & 1\end{pmatrix}$$

AR(1):

$$\boldsymbol{C}=\begin{pmatrix}1 & \rho & \rho^2 & \cdots & \rho^{T-1}\\ \rho & 1 & \rho & & \vdots\\ \rho^2 & \rho & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \rho\\ \rho^{T-.1} & \cdots & \rho^2 & \rho & 1\end{pmatrix}$$

CS:

$$\boldsymbol{C}=\begin{pmatrix}1 & \rho & \rho & \cdots & \rho\\ \rho & 1 & \rho & & \vdots\\ \rho & \rho & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \rho\\ \rho & \cdots & \rho & \rho & 1\end{pmatrix}$$

UN (UNR):

$$\boldsymbol{C}=\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1T} \\ \rho_{21} & 1 & \rho_{23} & & \vdots \\ \rho_{31} & \rho_{32} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \rho_{(T-1)T} \\ \rho_{T1} & \cdots & \rho_{T(T-2)} & \rho_{T(T-1)} & 1 \end{pmatrix}$$

CSH (an example for a heterogeneous model):

$$\boldsymbol{\Sigma}=\boldsymbol{D}^{1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}=\begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_T \end{pmatrix}\begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & & \vdots \\ \rho & \rho & \ddots & \ddots & \rho \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & \rho & 1 \end{pmatrix}\begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \sigma_T \end{pmatrix}$$

$$=\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \cdots & \rho\sigma_1\sigma_T \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & & \vdots \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \rho\sigma_T\sigma_{T-1} \\ \rho\sigma_T\sigma_1 & \cdots & \rho\sigma_T\sigma_{T-2} & \rho\sigma_T\sigma_{T-1} & \sigma_T^2 \end{pmatrix}$$

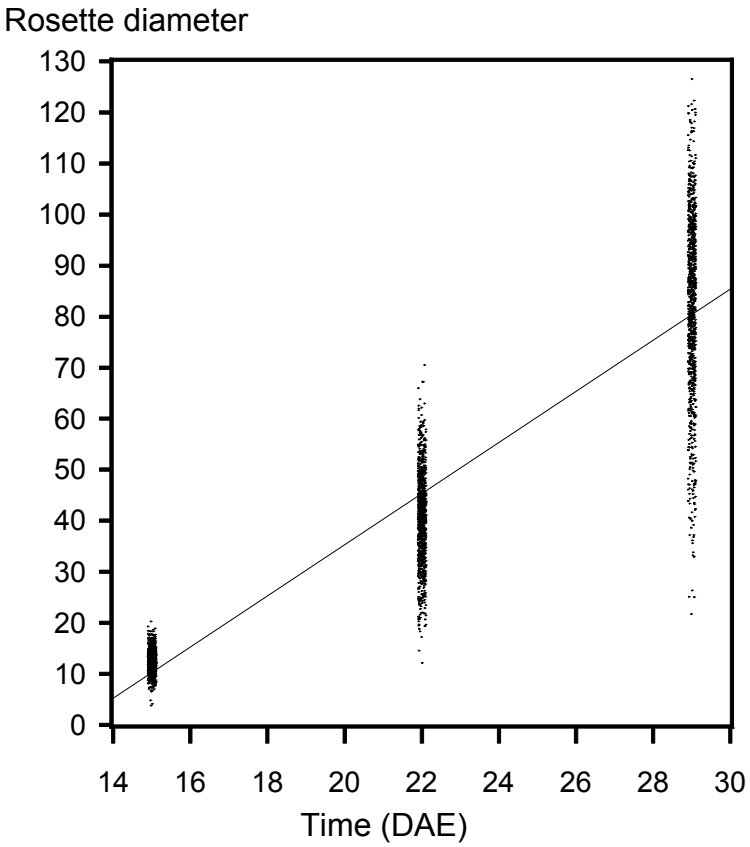


Fig. 9: Plot of rosette diameters against time for Arabidopsis data.

Fitting the same type of model simultaneously for all three structures, the following AIC values are found for different models:

Model[§]	AIC
Independent (ID)	13712.5
Independent heterogeneous (DIAG)	12024.2
Compound symmetry (CS)	13416.1
Heterogeneous compound symmetry (CSH)	11126.3
Autoregressive AR(1)	13198.3
Heterogeneous autoregressive ARH(1)	11050.0
Unstructured UN	11040.7

§ The same model was fitted simultaneously to all three structures $\text{var}(\boldsymbol{b}_{jk}) = \boldsymbol{\Sigma}_b$, $\text{var}(\boldsymbol{m}_{jkh}) = \boldsymbol{\Sigma}_m$, $\text{var}(\boldsymbol{s}_{jkhp}) = \boldsymbol{\Sigma}_s$.

The AIC values indicate that by far the dominating feature of the repeated variance-covariance structure is heterogeneity among time points. The unstructured model fits best and so is used for further analyses.

The unstructured model $\boldsymbol{\Sigma}_b$ for three time points can be represented as follows:

$$\boldsymbol{\Sigma_b} = \begin{pmatrix} \sigma^2_{1(b)} & \sigma_{1(b)}\sigma_{2(b)}\rho_{12(b)} & \sigma_{1(b)}\sigma_{3(b)}\rho_{13(b)} \\ \sigma_{1(b)}\sigma_{2(b)}\rho_{12(b)} & \sigma^2_{2(b)} & \sigma_{2(b)}\sigma_{3(b)}\rho_{23(b)} \\ \sigma_{1(b)}\sigma_{3(b)}\rho_{13(b)} & \sigma_{2(b)}\sigma_{3(b)}\rho_{23(b)} & \sigma^2_{3(b)} \end{pmatrix},$$

where $\sigma^2_{t(b)}$ is the the variance at the *t*-th time point and $\rho_{tt'(b)}$ is the correlation of the *t*-th and *t'*-th time point. The other two structures can be represented similarly. The variance-covariance estimates of the model are as follows:

Model term	$\boldsymbol{b}_{jk}$	$\boldsymbol{m}_{jkh}$	$\boldsymbol{s}_{jkhp}$
$\sigma^2_{1(b)}$	1.03	0.50	1.13
$\sigma^2_{2(b)}$	37.52	7.10	10.07
$\sigma^2_{3(b)}$	149.87	34.59	37.31
$\rho_{12(b)}$	0.9020	0.7047	0.6451
$\rho_{13(b)}$	0.8750	0.5202	0.5188
$\rho_{23(b)}$	0.9767	0.7012	0.7334

The variance is seen to increase drastically with time, and the correlation. Also, the correlation is smallest between the two most distant time points.

The fixed effects for factors of interest may be partitioned into main effects and interactions with time as usual. In particular, we will need a two-way structure for time × genotype, so we replace $\mu_t + \alpha_{it}$ by

$$\mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it},$$

where

μ = general mean
α_i = main effect for genotype
γ_t = main effect for time
$(\alpha\gamma)_{it}$ = time × genotype interaction

The full model can be written.

$$y_{ijkhpt} = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it} + r_{jt} + b_{jkt} + m_{jkht} + s_{jkhpt}$$

The resulting Wald-tests are as follows:

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
rep	2	1543	1.09	0.3378
rep*time	6	1543	237.30	<.0001
genotype	284	1543	8.05	<.0001
genotype*time	566	1543	4.78	<.0001

The model may be extended in various ways. For example, the structure of the genetic effects may be modelled by random effects (Section 4.8), allowing estimation of variance components for additive and non-additive effects (Kusterer et al. 2007). Time trends may be modelled by regressions as in the Duckweed example. These extensions are not elaborated here.

Instead of fitting a variance-covariance structure that allows for heterogeneity of variance, we may seek a simple data transformation that stabilizes variances. It turns out that a log-transformation stabilizes variances, but at the same time it introduces some nonlinearity (Fig. 10). For quantitative-genetic studies one may prefer to remain on the original scale (Kusterer et al., 2007), even though data transformation may seem desirable from a statistical point of view. It is then crucial to account for heterogeneity of variance.

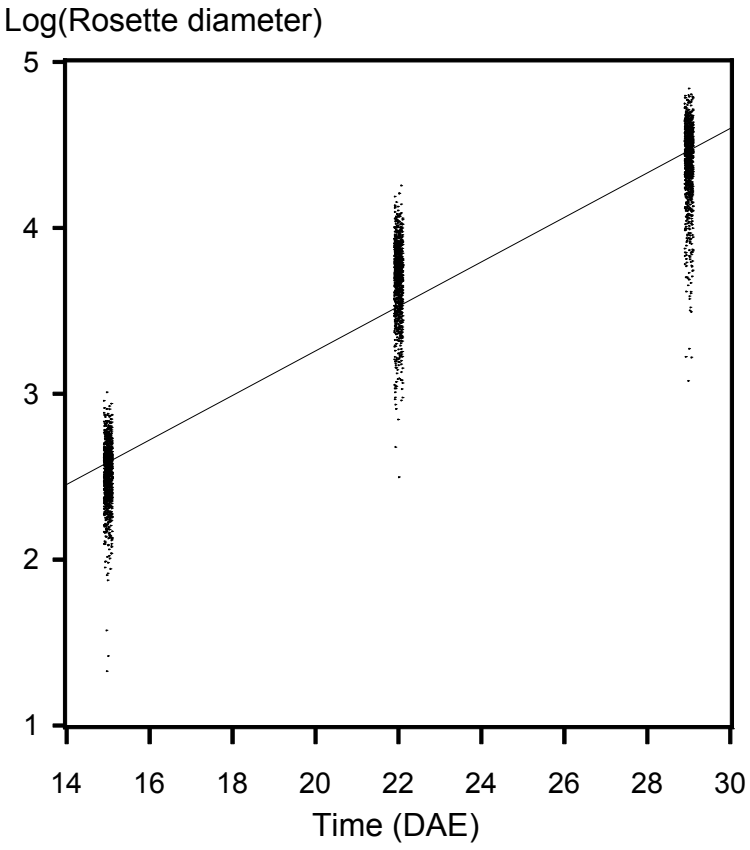


Fig. 10: Plot of logarithm of rosette diameters against time for Arabidopsis data.

Final remark: A strategy for setting up a mixed model for repeated measurements has been used here that is more fully described in Piepho et al. (2004).

7 Spatial models for field trials

Models for serial correlation among repeated measurements in short time series as discussed in Section 4.6 can also be used with benefit for the analysis of spatial correlation. Consider a single row of plots, as given in Fig. 11.

Fig. 11: A single row of plots. Arrow indicates direction of spatial correlation.

It is often reasonable to assume that neighboring plots are more similar than more distant plots. Thus, a variance-covariance model, in which correlation decays with spatial distance can be used. Essentially all time series models are suitable candidates and are, in fact, often used, for example AR(1) models.

There is one very common misconception regarding spatial analysis. The argument of some proponents of spatial analysis goes something like this: "The design was a randomized complete block design. Analysis of variance assumes that errors within blocks are uncorrelated. Inspection of real data reveals, however, that this assumption is incorrect, because neighbouring plots are positively correlated. Thus, it is incorrect to use analysis of variance, and it is better to use a spatial model." While it is true that a spatial model may be preferable because it can improve precision of the analysis by exploiting correlations, it is erroneous to assume that spatial correlation invalidates classical analysis of variance for blocked experiments, which uses a model with independent errors. Analysis of variance can be justified by randomization theory. This theory makes no assumptions whatsoever with respect to the correlation of plots within blocks. It merely assumes arbitrary plot values, which hold when all treatment have the same effect and considers the distribution of estimates of treatment effects and of F-statistics under a randomization distribution generated by random allocation of treatments to plots (Calinski and Kageyama, 2000). These results are generally valid, no matter how strongly correlated plots within blocks may be. Derivation from randomization theory in the end shows that data can be analysed as if errors were independent normal deviates with homogeneous variance. But this is just an outflow of randomization theory. Most importantly, this does not mean that for a fixed series of plots no spatial correlation must be present between plots for classical ANOVA to be valid. These issues may be somewhat difficult to fully appreciate without detailed study of randomization theory. The interested reader is encouraged to refer to standard texts on the subject, including Calinski and Kageyama (2000) and John and Williams (1995).

If field trials are to be analysed by spatial models, the same model selection task arises as in the analysis of repeated measures. It should be stressed that, as in the analysis of repeated measures, spatial models have no backing in randomization theory, so model choice is never clear-cut, and the danger of chosing the wrong model is always present. I here therefore advocate a somewhat conservative approach to spatial analysis that has been most prominently advocated by Williams (1986) and Williams et al. (2006). This conservative

approach rests on the fact that standard block analysis can be justified by randomization theory, if proper randomization has taken place. Thus, classical analysis by ANOVA with independent plot errors can always serve as a fall-back position, provided the experiment has been designed accordingly.

The approach emerging from this reasoning is to design experiments according to a standard blocking scheme, such as resolvable incomplete block or row-column designs, possibly optimized with some spatial analysis in mind. At the analysis stage, it may then be checked if a spatial variance-covariance component **within blocks** or **within rows and columns** can, in fact, improve upon a classical analysis. In this exploration, it is prudent to limit the set of models considered in order to avoid overfitting.

7.1 Revisiting the oats data of Section 3.8

Example 5 (continued): To exemplify the type of routine application advocate here, consider the results from a yield trial with oats laid out as an α-design reported in John and Williams (1995). This trial was already considered in Section 3.8 (p. 32), where a classical block analysis was provided. The trial had 24 genotypes, three complete replicates and six incomplete blocks within each replicate. The block size was four.

We may start analysis using the baseline model (see Secion 3.8)

$$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + e_{ijh}$$

where

y_{ijh} = yield of i-th genotype in h-th block nested within j-th complete replicate
μ = general effect
γ_j = effect of j-th complete replicate
b_{jh} = effect of h-th block nested within j-th complete replicate
τ_i = effect der i-th genotype
e_{ijh} = residual plot error associated with y_{ijh}

The base line model assumes independent error and block effects with

$$e_{ijh} \sim N\left(0, \sigma^2\right)$$
$$b_{jh} \sim N\left(0, \sigma_b^2\right)$$

This model yields an AIC of 68.9. We may then explore if addition of a spatial component improves the analysis. Specifically, we add a trend effect t_{ijh} for the plots:

$$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + t_{ijh} + e_{ijh}$$

We intentionally do not replace the residual independent error here, because in spatial analyses it often turns out that such an error term is needed in addition to a spatial component. Thus, plot error is decomposed into two components, one for trend (t_{ijh}) and one for residual measurement error (e_{ijh}). The measurement error is sometimes called **nugget effect** in geostatistical terminology.

We consider only two models for trend, mainly because blocks are rather small. Trend effects are assumed correlated for plots in the same block, while plots from different blocks are uncorrelated. Aside from the AR(1) model, we consider the **linear variance (LV)** model of Williams (1986), which states that the covariance decays linearly with spatial distance:

$$\text{cov}(t_1, t_2) = \sigma_t^2 (1 - \phi d),$$

where t_1 and t_2 are trend values on two plots and d is the distance of these two plots. By contrast, using the same formulation, the AR(1) model has

$$\text{cov}(t_1, t_2) = \sigma_t^2 \rho^d \quad (0 < \rho < 1)$$

Note that in both cases, when $d = 0$ we have the variance of a trend effect equal to σ_t^2. For trend, one typically assumes that generally

$$\text{var}(t_1) = \text{var}(t_2) = \sigma_t^2 \quad \text{and}$$
$$\text{cov}(t_1, t_2) = \sigma_t^2 f(d)$$

where d is the distance between the two plots and $f(d)$ is some smooth decreasing function of d with $f(0) = 1$.

The following AIC values are found for different models (random block effects fitted throughout):

Model for plot effects		AIC	Mean variance of a difference
e_{ijh}	t_{ijh}		
ID	-	68.9	0.070
-	LV	66.4	0.056
ID	LV	68.3	0.059
-	AR(1)	66.4	0.061
ID	AR(1)	§68.9	0.070

§ converged to baseline model

Both the LV and AR(1) model for trend seem to do better than the baseline model. It is noteworthy that the residual error e_{ijh} does not appear necessary with a spatial component added, so we may drop this component. The LV and AR(1) models for trend without residual error have comparable AIC. LV has the smaller variance of a difference and so could be preferred for final analysis.

7.2 Model diagnosis by semivariogram

We have seen that likelihood-based methods such as the LR test or AIC can be used conveniently to select a model for the variance-covariance structure. As opposed to repeated measures data, it is usually impossible to fit an unstructured model due to lack of data, so it is

not possible to formally test for lack-of-fit. Even if a model has the most favorable AIC value, it is difficult to tell whether the selected model is really fully appropriate. Careful analysis will always inspect residuals for any lack-of-fit. For spatial data, a particularly useful display is the so-called semi-variogram. It derives its name from the way the ordinate values are computed: In the absence of and treatment of block effects, the observed semivariance is just the halved squared difference of two observations:

$$v = \frac{1}{2}(y_1 - y_2)^2 \ .$$

Assuming there are no treatment and block effects, the model for *y* can be written as

$$y = \mu + t + e$$

where *t* is spatial trend, following some spatial process such as AR(1), and *e* is independent residual error. With this model the expected value of *v* can be written as

$$E(v) = \frac{1}{2}E\left[(y_1 - y_2)^2\right] = \frac{1}{2}E\left[(t_1 + e_1 - t_2 - e_2)^2\right]$$
$$= \frac{1}{2}\left[E(t_1^2) - 2E(t_1 t_2) + E(t_2^2)\right] + \frac{1}{2}\left[E(e_1^2) - 2E(e_1 e_2) + E(e_2^2)\right]$$
$$= \frac{1}{2}\left[\mathrm{var}(t_1) - 2\,\mathrm{cov}(t_1, t_2) + \mathrm{var}(t_2)\right] + \frac{1}{2}\left[\mathrm{var}(e_1) - 2\,\mathrm{cov}(e_1, e_2) + \mathrm{var}(e_2)\right]$$

Now measurement errors are assumed to be independent with constant variance σ^2, so we have

$$\mathrm{var}(e_1) = \mathrm{var}(e_2) = \sigma^2 \quad \text{and}$$
$$\mathrm{cov}(e_1, e_2) = 0 \ .$$

For trend, one typically assumes that

$$\mathrm{var}(t_1) = \mathrm{var}(t_2) = \sigma_t^2 \quad \text{and}$$
$$\mathrm{cov}(t_1, t_2) = \sigma_t^2 f(d) \ ,$$

where *d* is the distance between the two plots (see Section 7.1). With these assumptions the observed semivariance has the following expectation:

$$E(v) = \gamma(d) = \frac{1}{2}\left[\sigma_t^2 - 2\sigma_t^2 f(d) + \sigma_t^2\right] + \frac{1}{2}\left[\sigma^2 + \sigma^2\right] = \sigma^2 + \sigma_t^2\left[1 - f(d)\right]$$

For example, under the linear variance model we have $f(d) = 1 - \phi d$ and thus

$$E(v) = \gamma(d) = \sigma^2 + \sigma_t^2\left[1 - f(d)\right] = \sigma^2 + \sigma_t^2 \phi d \ ,$$

so the semi-variogramm is a linearly increasing function with intercept σ^2 and slope $\sigma_t^2 \phi$. Under the AR(1) model we have $f(d) = \rho^d$ and thus

$$E(v) = \gamma(d) = \sigma^2 + \sigma_t^2[1 - f(d)] = \sigma^2 + \sigma_t^2\left(1 - \rho^d\right) .$$

This function approaches the value $\sigma^2 + \sigma_t^2$ as distance *d* increases, because then ρ^d tends to zero. This limiting value is also known as **sill**. The distance at which the sill is approached is known as the **range**. The intercept of the semi-variogram is equal to σ^2, which is also known as the **nugget**. There are many other nonlinear models for the semivariogramm, which smoothly approach a sill, including the Gaussian model and the Spherical model. We do not give details here, but depict these models along with the LV and AR(1) models in Fig. 12.

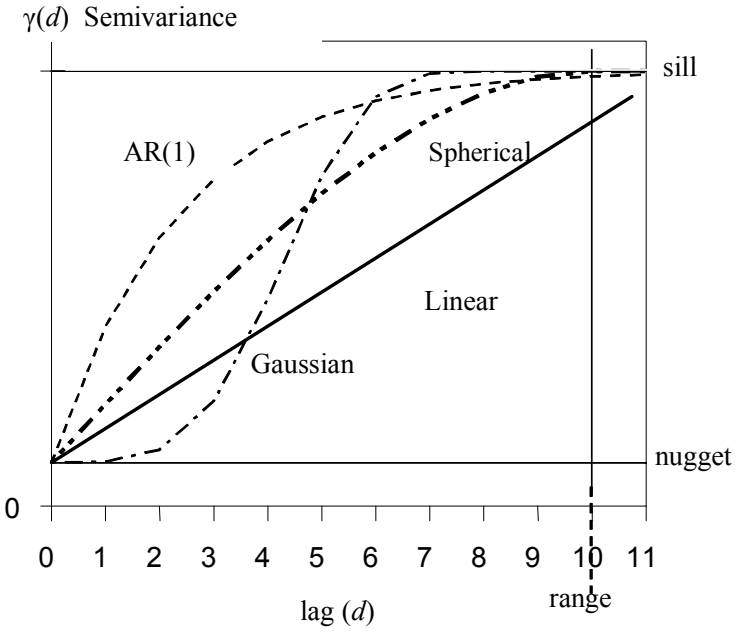


Fig. 12: Examples of some semi-variogram models, corresponding to spatial covariance models.

So far we have assumed for simplicity that data are free of block and treatment effects. In most experiments, such effects are present. In this case, effects are removed by computing residuals based on a linear model fit with treatment and block effects.

7.3 The wheat data of Besag and Kempton

Example 14: In order to further exemplify spatial models, we analyse the yield data of Table 1 in Besag and Kempton (1986) (**besag & kempton.dat**). Plots were arranged in seven columns of 52 plots each. There were two wheat cultivars grown alternatingly according to a chessboard pattern, so that each plot had four neighbouring plots with the opposing cultivar. Spatial covariance was modelled along columns. A robust semi-variogram was computed down each column, using residuals from a linear model with fixed main effects for cultivar and column, and averaged this across columns (see Fig. 13; Piepho, 2008b). It is seen that either a linear or a simple non-linear model such as AR(1) seem to hold well up to lags (plot distances *d*) of about 20. Apparently, it is difficult to find a simple model that properly describes the spatial correlation structure. The figure also shows that a nugget (measurement error *e*) is needed to model trend, because the semi-variogram has a non-zero intercept.

The reasons for this failure of simple models may be manifold. For example, there may be strong global trend. There are many sophisticated methods for modelling trend in such cases, including the addition of random effects for technical errors (tractor wheeling etc.) and fitting of splines for trend. It is not usually a simple task to identify a suitable model.

A simpler route of attack is to fit spatial models only up to short distances. If the design involved incomplete blocks of size between about 5 and 10, it is rather more likely that one of the common simple spatial models will hold within blocks. The semivariogram for the wheat data indicates that for blocks of size about 7 a LV or AR(1) model will hold well.

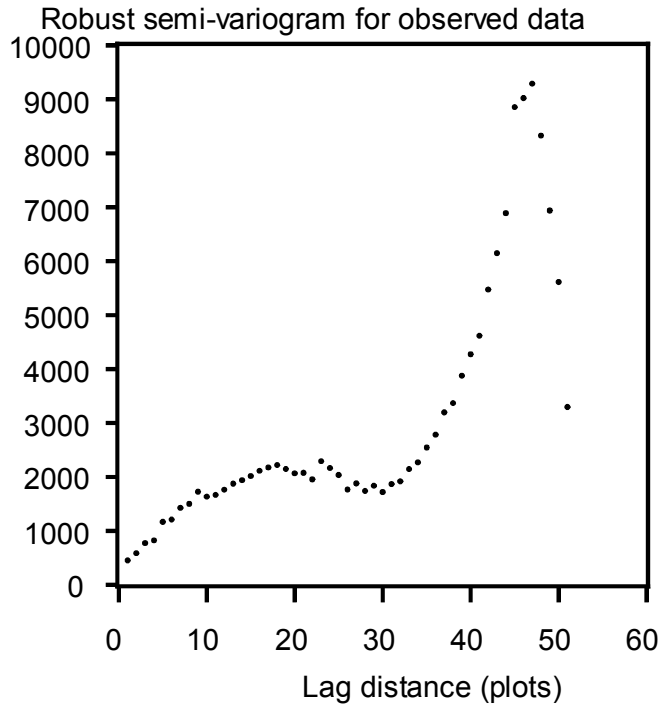


Fig 13: Robust semi-variogram for wheat data of Besag & Kempton (1986).

To explore this hypothesis, I divided columns into eight blocks, with block sizes alternating between six and seven. The resulting analysis is shown in Table 10. It is seen that the linear variance (LV) model marginally beats the AR(1) model in terms of AIC, when a nugget is fitted, agreeing well with the semi-variograms in Fig. 1. Presence of a nugget variance in the model has quite a notable effect on the standard error of the contrast between the two genotypes, while the contrast estimate itself is little affected.

Table 10: Restricted maximum likelihood (REML) fits for wheat data of Besag and Kempton (1986). Columns subdivided into eight blocks alternating in size between six and seven. Blocks fitted as random effect.

Variance-covariance model	$-2 \log L_R$	AIC	$\hat{\beta}_1 - \hat{\beta}_2$	s.e.$(\hat{\beta}_1 - \hat{\beta}_2)$	$^{\S}\hat{\rho}$
Original data - Random blocks (7 columns, 8 blocks each):					
Linear variance	3553.5	3557.5	-36.75	1.7684	-
AR(1)	3514.7	3520.7	-36.71	2.0697	0.4122
Linear variance + nugget	3508.5	3514.5	-36.66	2.2248	-
AR(1) + nugget	3508.6	3514.6	-36.67	2.2044	0.9069

§ Autocorrelation coefficient of AR(1) model.

4.7.4 Extension in two dimensions

It is possible to extend the spatial model in two dimensions along rows and columns. The simplest way to extend a spatial model in two dimensions is to use so-called direct product (Kronecker product) structures with one term for rows and one for columns. For details see e.g. Gilmour et al. (1997) and Williams et al. (2006).

The most common direct product structure is an AR(1) x AR(1) model, which assumes that an AR(1) model holds both along rows and along columns. Rows and columns have their own model. To come to an overall model, the correlation matrices for rows and columns, given by

$$\boldsymbol{\Sigma_r} = \begin{pmatrix} 1 & \rho_r & \rho_r^2 & \dots & \rho_r^{c-1} \\ \rho_r & 1 & \rho_r & \dots & \rho_r^{c-2} \\ \rho_r^2 & \rho_r & 1 & \dots & \rho_r^{c-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_r^{c-1} & \rho_r^{c-2} & \rho_r^{c-3} & \dots & 1 \end{pmatrix} \text{ for correlation in rows and}$$

$$\boldsymbol{\Sigma_c} = \begin{pmatrix} 1 & \rho_c & \rho_c^2 & \dots & \rho_c^{r-1} \\ \rho_c & 1 & \rho_c & \dots & \rho_c^{r-2} \\ \rho_c^2 & \rho_c & 1 & \dots & \rho_c^{r-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_c^{r-1} & \rho_c^{r-2} & \rho_c^{r-3} & \dots & 1 \end{pmatrix} \text{ for correlation columns ,}$$

are connected by a direct product as

$$\boldsymbol{V} = \sigma_{rc}^2 \boldsymbol{\Sigma_r} \otimes \boldsymbol{\Sigma_c} \quad .$$

The structure has three parameters: the spatial variance σ_{rc}^2 and the autocorrelations for rows and columns, ρ_r and ρ_c. To illustrate the direct product operator, consider the simple case of just two rows and two columns, where

$\boldsymbol{\Sigma}_r = \begin{pmatrix} 1 & \rho_r \\ \rho_r & 1 \end{pmatrix}$ and $\boldsymbol{\Sigma}_r = \begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix}$, whence

$$\boldsymbol{V} = \sigma^2_{rc}\boldsymbol{\Sigma}_r \otimes \boldsymbol{\Sigma}_c = \sigma^2_{rc}\begin{pmatrix} 1\begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix} & \rho_r\begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix} \\ \rho_r\begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix} & 1\begin{pmatrix} 1 & \rho_c \\ \rho_c & 1 \end{pmatrix} \end{pmatrix} = \sigma^2_{rc}\begin{pmatrix} 1 & \rho_c & \rho_r & \rho_r\rho_c \\ \rho_c & 1 & \rho_c & \rho_r \\ \rho_r & \rho_r\rho_c & 1 & \rho_c \\ \rho_r\rho_c & \rho_r & \rho_c & 1 \end{pmatrix}$$

Plots in the same row, but one column apart have correlation ρ_r.
Plots in the same column, but one row apart have correlation ρ_c.
Plots one row and one column apart have correlation $\rho_r\rho_c < \rho_r$ or ρ_c.

Generally, the correlation is $\rho_r^{d_r}\rho_c^{d_c}$, where d_r and d_c are the distances (in numbers of plots) along rows and columns, respectively.

Example 7 (continued): Different spatial models were fitted to the data of the row-column design considered in Section 3.10 (p.34; **rowcol.dat**). In addition to the two-dimensional model, we also fitted the AR(1) along rows only and along columns only. These models correspond to the choices

$\boldsymbol{\Sigma}_r = \boldsymbol{I}_r$ or
$\boldsymbol{\Sigma}_c = \boldsymbol{I}_c$.

For example, when $\boldsymbol{\Sigma}_r = \boldsymbol{I}_r$, then

$$\boldsymbol{V} = \sigma^2_{rc}\boldsymbol{I}_r \otimes \boldsymbol{\Sigma}_c = \sigma^2_{rc}\begin{pmatrix} \boldsymbol{\Sigma}_c & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_c & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_c & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_c \end{pmatrix},$$

so rows become independent subjects.

Spatial correlation was assumed to extend only across a replicate. For each model we tried adding a nugget effect. The spatial models with nuggets caused convergence problems, with the nugget approaching a very tiny value. The AR(1) x AR(1) model has a favourable AIC value, but it is outperformed by an AR(1) model along columns only, while the AR(1) model along rows only showed no improvement over baseline. It is tempting here to use the 2-dimensional AR(1) x AR(1) model, because it produces the smallest mean s.e.d., but this would likely be too optimistic because AIC indicates otherwise.

Model	AIC	mean s.e.d.
Baseline (random rows and columns)	76.0	0.3853670
AR(1) x AR(1)	74.3	0.2914215
AR(1) x AR(1) + nugget	§	§
AR(1) along rows only	77.8	0.3837451
AR(1) along rows only + nugget	79.8	0.3837476
AR(1) along columns only	72.4	0.3227287
AR(1) along columns only + nugget	----- did not converge -----	

§ converges to AR(1) x AR(1) without nugget

8 Random genotype effects

8.1 Selection and shrinkage

It is a common experience of plant breeders that the progeny of selected genotypes tend to fall back or shrink towards the mean. One of the earliest accounts of this common experience is the famous selection experiment by the Danish geneticist JOHANNSEN, who selected seeds from a population of Princess beans (Fig. 14). To his surprise he found that the progeny from big seeds had the same size distribution as had the progeny from small seeds. The explanation for this observation is that Princess beans are autogamous and that the parent population was genetically completely homogeneous, i.e., each individual had the same genotype. Also, due to self-pollination, all individuals were homozygous. Thus, all variance in seed size observed in the parent population of beans was entirely due to environmental effects. Consequently, the response to selection was zero.

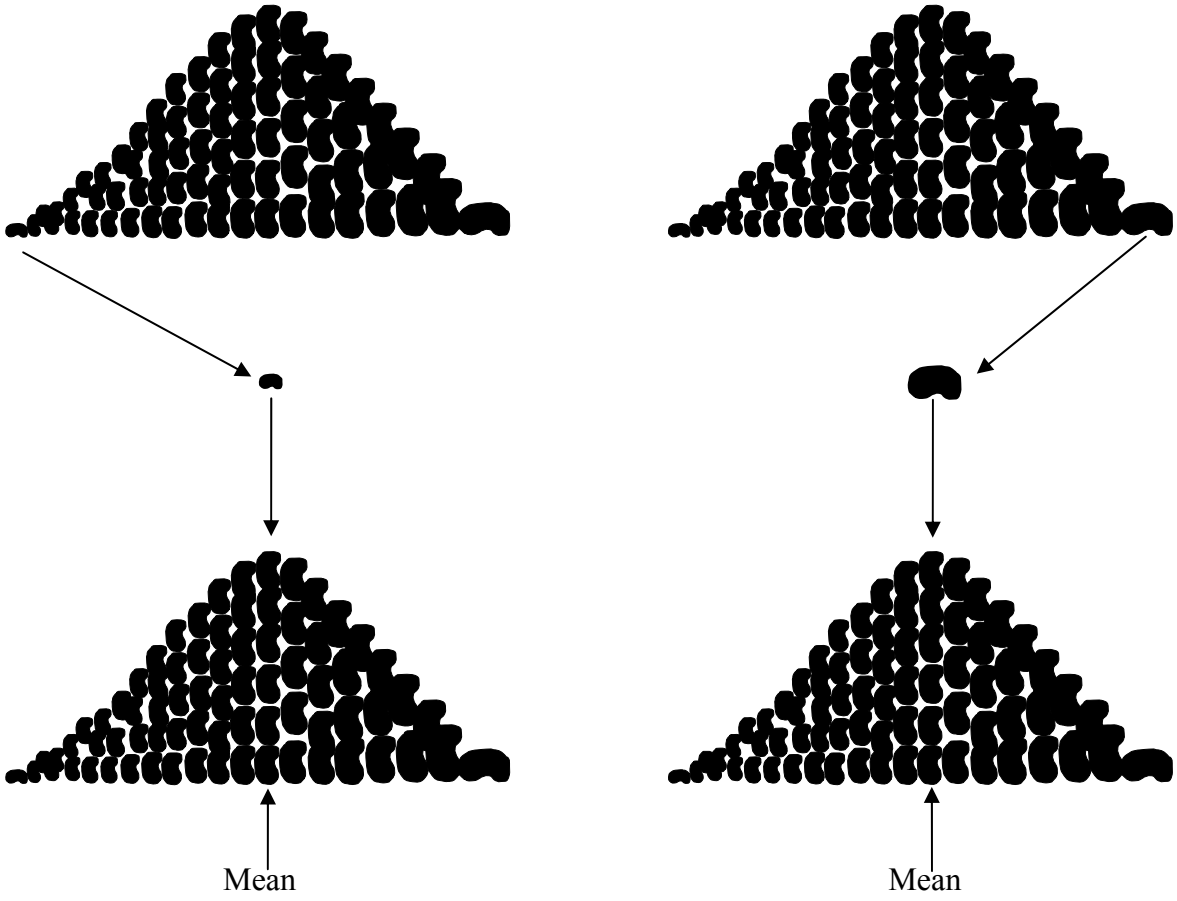


Fig. 14: Progeny of selection from a genetically homogeneous line of Princess beans (*Phaseolus vulgaris* L.). This is the famous selection experiment by the Danish geneticist JOHANNSEN conducted in 1903. No matter whether we select small seeds (left) or big seeds (right), the progeny have the same size distribution, because all variation is caused by the environment and there is **no genetic variance**—the mean of the progeny is shrunken to the mean of the parental population ($h^2 = 0$).

If there were no environmental effects, then response to selection for an autogamous crop would be 100%, providing there is any genetic variance. This hypothetical scenario is depicted in Fig. 15. It is now assumed that all individuals are genetically distinct and that they

are homozygous at all loci. It is further assumed that there are no environmental effects. Since all variance in the parent population is genetically determined, there is no shrinkage of the selected seeds towards the mean of the parent population.

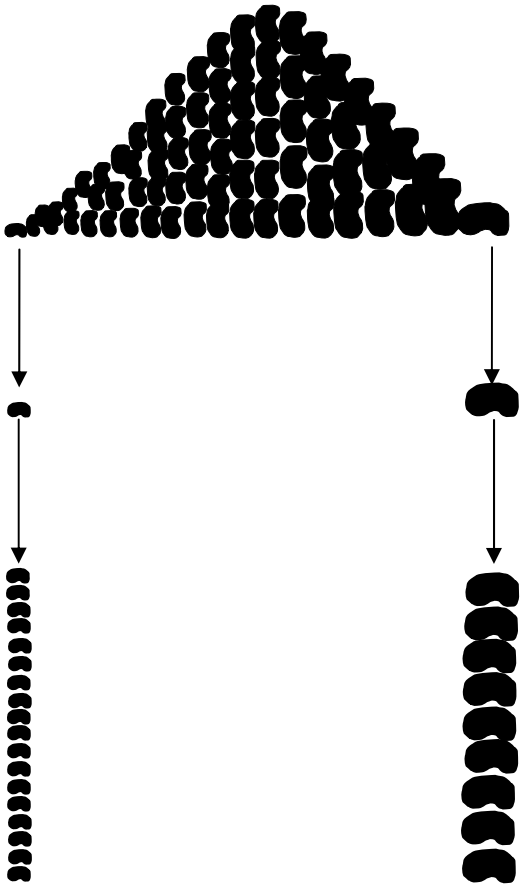


Fig. 15: Progeny of selection from a genetically heterogeneous population of *Miracle Beans* (hypothetical!): There is no **environmental variance**, and the response to selection is perfect—there is no shrinkage of the progeny towards the mean of the parental population ($h^2 = 1$).

The hypothetical example with no environmental variance and the real example with no genetic variance are the two extremes: No shrinkage is observed for the selected progeny in the former case and complete shrinkage is found in the latter. Often, the response to selection is intermediate between these two extremes, so the observed shrinkage towards the mean is only partial. This will be the case whenever there is both genetic and environmental variance.

8.2 Best linear unbiased prediction of genetic effects (BLUP)

BLUP is a method of estimation for genetic effects, which in a sense tries to anticipate the shrinkage towards the mean observed in the progeny. Instead of waiting for that disappointing outcome in the next season, BLUP computes a shrunken estimate of genetic effects in the current season. Thus, genotypes with a large least squares mean will be shrunken downwards, while genotypes with small least squares means will be shrunken upwards. The degree of shrinkage will depend on both the genetic and the environmental variation as assessed by variance components.

BLUP is effected in a mixed model analysis by taking genotype effects as random. The random assumption may be justified by the fact that in breeding programs, genotypes are usually progeny in a segregating population. The genotypes actually observed may be regarded as a random sample from a hypothetical population comprising all potential genotypes that might have been observed according to Mendelian laws of inheritance. Based on this premise, it is reasonable to consider genotypes as random. This assumption is the basis of a huge body of quantitative-genetic theory, including selection theory (Bos und Caligari, 1995).

Estimation of effects under the random effects assumption (BLUP) is conceptually different from estimation under a fixed effects assumption (best linear unbiased estimation/estimator – BLUE). BLUE was the method of estimation in the preceding section: Least squares means are BLUE of the genotype means. The transition from BLUE to BLUP is a trivial matter, in so far as implementation with a mixed model package is concerned: just declare the genotypic effect τ_i as a random effect instead of as a fixed effect. The purpose of this section is to explain the central statistical idea underlying BLUP, using a simple example.

Example 15: In a small breeding program with red radish, 26 strains were tested in a randomised complete block design with two blocks. (S. Schmiegel, Universität-Gesamthochschule Kassel, pers. communication, 2001). Among other traits, a sample of plants per plot was scored for woodiness. Based on these data, the percentage of woody plants can be determined (**radieschen.dat**) as $y = z/m$., were m is the total number of plants and z is the number of woody plants.

		No. of plants				No. of plants	
Block	Line	Woody z	Total m	Block	Line	Woody z	Total m
1	1	18	41	1	14	12	44
2	1	19	41	2	14	25	50
1	2	18	35	1	15	12	40
2	2	19	37	2	15	18	37
1	3	22	42	1	16	8	44
2	3	28	47	2	16	16	45
1	4	9	32	1	17	12	47
2	4	12	41	2	17	12	45
1	5	13	45	1	18	14	44
2	5	14	44	2	18	21	38
1	6	6	46	1	19	24	50
2	6	8	40	2	19	27	45
1	7	11	38	1	20	7	43
2	7	8	39	2	20	5	39
1	8	8	37	1	21	33	45
2	8	17	41	2	21	19	32
1	9	2	40	1	22	12	44
2	9	5	45	2	22	24	40
1	10	10	37	1	23	17	36
2	10	7	46	2	23	29	42
1	11	21	37	1	24	6	43

2	11	17	33	2	24	17	47
1	12	3	34	1	25	7	33
2	12	8	40	2	25	8	36
1	13	19	38	1	26	5	40
2	13	18	44	2	26	12	45

Our objective is to estimate genetic effects based on the assumption of random genotypes. As usual, we use the model

$$y_{ij} = \mu + b_j + \tau_i + e_{ij} \quad ,$$

where y_{ij} is the observed value of the *i*-th genotype in the *j*-th block, b_j is the effect of the *j*-th block, τ_i is the effect of the *i*-th genotype, and e_{ij} is the residual effect corresponding to y_{ij}. Now, genotypic effects τ_i are assumed to be random with zero mean and variance σ_τ^2. The error e_{ij} is random with zero mean and variance σ_e^2. Our objective is to estimate genotype means, which may be expressed as

$$\eta_i = \mu + \frac{b_1 + b_2}{2} + \tau_i \quad .$$

If genotype effects were taken as fixed, out best estimator would be the simple genotype mean

$$BLUE(\eta_i) = \bar{y}_{i\bullet} .$$

This is a good estimator in the sense that it has the smallest variance compared to alternative estimators and "in the long run" the estimator will yield the correct value η_i. More formally, we may say that the estimator is unbiased, i.e., its expected value equals η_i, i.e.,

$$E(\bar{y}_{i\bullet}) = \eta_i \quad ,$$

where E(.) denotes the expected value.

Now alluding to the bean example, we may contemplate an alternative simple estimator of genotype means: If there is no genetic variance, and all variance is environmentally determined, then all η_i are the same and we expect all progeny to be shrunken entirely to the overall mean. Therefore, in the extreme case where the genetic variance is zero, we would use the overall sample mean $\bar{y}_{\bullet\bullet}$ as the best estimator of η_i, thus anticipating the shrinkage to be expected in the progeny. In actuality, complete shrinkage will not occur in breeding programs, were pedigrees are designed to create a maximum of genetic variance! But still, we have both genetic and environmental effects, so some shrinkage is expected in the progeny, and it is useful to consider an estimator that anticipates this shrinkage.

To summarize, we have considered two alternative estimators: When all variance is genetically determined, then the genotype mean $\bar{y}_{i\bullet}$ is the best estimator. When all variance is environmentally controlled, then the overall mean $\bar{y}_{\bullet\bullet}$ is the best estimator. In breeding trials there will usually be both genetic and environmental variance. Thus, we may contemplate an estimator, which provides a compromise between the simple genotype mean $\bar{y}_{i\bullet}$ and the

overall mean $\bar{y}_{\bullet\bullet}$:

$$\tilde{\eta}_i = w\bar{y}_{i\bullet} + (1-w)\bar{y}_{\bullet\bullet} \quad .$$

The tilda (~) on top of the mean η_i indicates that we are estimating η_i. The weights w and $(1-w)$ in this estimator sum to zero. This restriction is imposed to make sure that the estimator is unbiased in the sense that its expected value over a random sample of genotypes equals the expected value of η_i over a random sample of genotypes:

$$E(\tilde{\eta}_i) = E(\eta_i) = \mu + \frac{b_1 + b_2}{2} .$$

This notion of unbiasedness is different from that of BLUE, where $E(\bar{y}_{i\bullet}) = \eta_i$. Clearly, for a given genotype our alternative estimator $\tilde{\eta}_i$ is a biased estimator of η_i, i.e. its expected value does not equal η_i, and thus

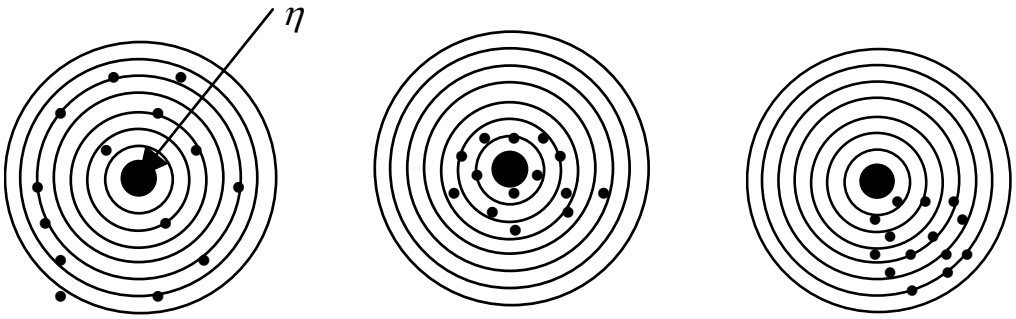
$$Bias(\tilde{\eta}_i) = E(\tilde{\eta}_i) - \eta_i \neq 0 \quad .$$

The bias comes from the shrinkage property of the estimator. Before discussing the specific choice of w, we will consider the important question **why it can be useful to accept an estimator which is biased**.

A more comprehensive measure of accuracy than the bias of an estimator is the mean squared error of estimation (MSE). In the present context this is defined as the expected value of the squared difference $(\tilde{\eta}_i - \eta_i)^2$, where the expectation is taken across the whole population of genotypes. A good estimator should have a small MSE. Bias is only one component of the MSE. The other component is the variance of estimation. We may write

$$MSE(\tilde{\eta}_i) = E\left[(\tilde{\eta}_i - \eta_i)^2\right] = Variance(\tilde{\eta}_i) + [Bias(\tilde{\eta}_i)]^2 \quad .$$

Now some bias in an estimator can be tolerated if its variance is much smaller than that of alternative, possibly unbiased estimators. This important fact will now be illustrated using an analogy: that of a target and bullets shot from a pistol (Fig. 16).



Large variance,
no bias
⇒ large MSE

Small variance,
no bias
⇒ small MSE

Small variance,
with bias
⇒ large MSE

Fig. 16: Bias and variance of estimates illustrated with a target and bullets shot from a pistol.

The center of the target corresponds to the quanity to be estimated. In our case, this is the genotype mean η_i. A shot on the target corresponds to an estimate of η_i in an experiment. If we imaging several repetitions of the same experiment, we can think of each experiment as being represented by a shot from our pistol.

For the left target the pistol is not accurate, because the shots scatter quite widely. If we measured the distance from the center, we would find a large variance. There is no bias, however, because "on average" or "in the long run" we do hit the center of target. If we measured the coordinates of the shots and took the average across shots, we would be quite close to the center of the target. We here have a picture for an unbiased estimator. The example shows that an unbiased estimator is not necessarily a good estimator, i.e., when the variance is large. Even though the bias is zero, the MSE is large.

For the target in the middle, the variance is smaller than for the left target, because the points do not scatter quite as much. This picture corresponds to an unbiased estimator, which is also an accurate estimator, because it has a small variance; its MSE is therefore smaller than that for the left target.

The right target shows the same variance as the target in the middle, but the variation is not around the center of the target. Clearly, the pistol is biased in this case. This corresponds to a biased estimator, which has a relatively small variance. Its MSE will be about the same as for the left target.

The target analogy shows that some bias can be accepted, providing the MSE is smaller than that of other estimators. When using BLUP, we accept some bias, trading this for a smaller variance. The net effect of the transition from BLUE to BLUP is a reduction in MSE.

In our case, BLUP corresponds to use of the weighted estimator $\tilde{\eta}_i = w\bar{y}_{i\bullet} + (1-w)\bar{y}_{\bullet\bullet}$. It turns out that the MSE of this estimator is minimized by using the weight

$$w = h^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \frac{\sigma_e^2}{r}} = \frac{\mathrm{var}(\tau_i)}{\mathrm{var}(\bar{y}_{i\bullet})} \quad ,$$

where σ_τ^2 and σ_e^2 are the genetic variance and the error variance, respectively, and r is the number of replicates per genotype. This weight is also known as the broad-sense heritability (h^2). Note that h^2 must take on a value between zero and unity, i.e., $0 \le h^2 \le 1$. The heritability gives the fraction of the variance in the simple mean $\bar{y}_{i\bullet}$ that is genetically determined. In other words, h^2 is the ratio of genetic variance, i.e., the variance of genetic effects τ_i, and phenotypic variance, i.e., the variance of the sample means $\bar{y}_{i\bullet}$. To highlight this important fact, we give the equation in prose form:

$$h^2 = \frac{\text{genetic variance}}{\text{phenotypic variance}} \quad .$$

If the genetic variance is zero, and all phenotypic variance is environmentally determined, then $h^2 = 0$. If the environmental variance is zero, and all phenotypic variance is genetically determined, then $h^2 = 1$.

Now consider our weighted estimator $\tilde{\eta}_i = w\bar{y}_{i\bullet} + (1-w)\bar{y}_{\bullet\bullet}$, which may be re-written as $\tilde{\eta}_i = h^2\bar{y}_{i\bullet} + (1-h^2)\bar{y}_{\bullet\bullet}$. When $h^2 = 0$, all weight is given to the general mean $\bar{y}_{\bullet\bullet}$, while when $h_2 = 1$, all weight is given to the genotype mean $\bar{y}_{i\bullet}$.

Our weighted estimator of η_i can also be written as

$$BLUP(\eta_i) = \tilde{\eta}_i = \bar{y}_{\bullet\bullet} + h^2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \quad ,$$

and the corresponding estimator of the genetic effect can be written as

$$BLUP(\tau_i) = \tilde{\tau}_i = h^2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \quad . \tag{1}$$

This expression nicely shows the shrinkage property of the estimator. When $h^2 = 0$, then $\tilde{\eta}_i = \bar{y}_{\bullet\bullet}$ and $\tilde{\tau}_i = 0$, i.e., we have complete shrinkage to the mean. When $h^2 = 1$, then $\tilde{\eta}_i = \bar{y}_{i\bullet}$ and $\tilde{\tau}_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$, i.e., we have no shrinkage to the mean.

In this context it is perhaps useful to re-iterate that heritability may be used to estimate the response to selection. We may write (Bos und Caligari, 1995):

$$R = h^2 S \quad ,$$

where S is the phenotypic mean deviation of the selected genotypes from the mean (selection differential) and R is the response to selection defined as the mean deviation of the progeny from the overall mean. When $h^2 = 0$, then $R = 0$, i.e. the response to selection will be zero (see JOHANNSEN's experiment!).

In order to use BLUP in practice, we need to use estimates of variance components. Using estimates in place of known parameters reduces the efficiency somewhat.

Example 15 (continued): For the radish data, BLUE and BLUP are shown in Table 11 and in Fig. 17. We can clearly see the shrinkage toward the mean.

Table 11: BLUE and BLUP for the radish data.

Strain	$BLUP(\eta_i)$	$BLUE(\eta_i)$	$BLUP(\tau_i) = \hat{h}^2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$	$BLUE(\tau_i) = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$
1	0.4353	0.4512	0.08725	0.10313
2	0.4884	0.5139	0.1403	0.16581
3	0.5272	0.5598	0.1791	0.21169
4	0.2964	0.2870	-0.05171	-0.06112
5	0.3104	0.3035	-0.03769	-0.04455
6	0.1934	0.1652	-0.1547	-0.18287
7	0.2628	0.2473	-0.08526	-0.10078
8	0.3205	0.3154	-0.02763	-0.03266
9	0.1218	0.0806	-0.2263	-0.26753
10	0.2323	0.2112	-0.1158	-0.13686
11	0.5116	0.5414	0.1635	0.19327
12	0.1755	0.1441	-0.1726	-0.20397

13	0.4382	0.4545	0.09007	0.10646
14	0.3805	0.3864	0.03238	0.03828
15	0.3863	0.3932	0.03820	0.04516
16	0.2809	0.2687	-0.06717	-0.07940
17	0.2744	0.2610	-0.07368	-0.08709
18	0.4220	0.4354	0.07387	0.08732
19	0.5104	0.5400	0.1624	0.19191
20	0.1767	0.1455	-0.1714	-0.20259
21	0.6150	0.6635	0.2669	0.31546
22	0.4228	0.4364	0.07468	0.08828
23	0.5454	0.5813	0.1973	0.23326
24	0.2656	0.2506	-0.08246	-0.09747
25	0.2373	0.2172	-0.1108	-0.13091
26	0.2193	0.1958	-0.1288	-0.15225

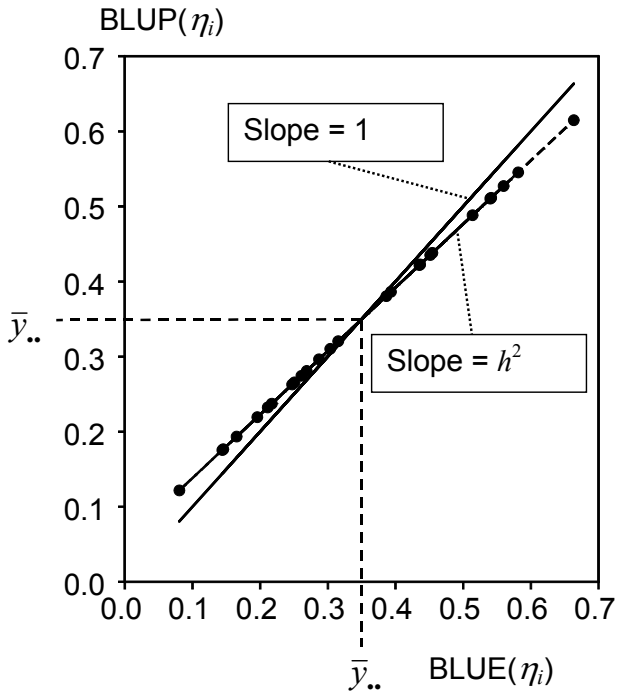


Fig. 17: BLUP versus BLUE for 26 radish strains. $h^2 = 0.846$.

The variance component estimates are:

$\hat{\sigma}_\tau^2 = 0.021263$

$\hat{\sigma}_e^2 = 0.00774$

The resulting estimate of h^2 is:

$$\hat{h}^2 = \frac{\hat{\sigma}_\tau^2}{\hat{\sigma}_\tau^2 + \frac{\hat{\sigma}_e^2}{r}} = \frac{0.021263}{0.021263 + \frac{0.00774}{2}} = 0.846$$

In the present example, shrinkage was the same for all genotypes, because the design was balanced. Thus, ranking of genotypes is identical for BLUE and BLUP, and it does not practically matter whether we use BLUE or BLUP for selection. There are, however, two

important reasons to favour BLUP over BLUE in more complex applications.

Incomplete blocking: With incomplete blocking, BLUE for genotype means differ in accuracy, and therefore shrinkage differs among genotypes. Thus, rankings may differ between BLUE and BLUP. It has been shown (Searle et al., 1992) that BLUP maximizes the probability of identifying the true ranking of η_i.

Genetic covariance: Often, the genotypes under test are related through a more or less complex pedigree. The pedigree induces genetic covariances among related progeny. The genetic covariance can also be incorporated into a mixed model, as we will see later, and BLUP based on such a model is typically more efficient than BLUE (Searle et al., 1992; Panter and Allan, 1995).

8.3 Genetic covariance in pedigrees

When using BLUP, we are regarding genotype effects as random. In breeding applications, genotypes are usually related due to the pedigree. The relatedness induces **genetic correlation** or **genetic covariance**. The closer the relatedness, the higher the genetic correlation. For example, assuming absence of epistasis and dominance, the genetic correlation of two inbred lines equals

$a_{XY} = 2f_{XY}$,

where f_{XY} is the **coefficient of coancestry** between inbred lines X and Y. The genetic covariance of two individuals is equal to

$$a_{XY}\sigma_u^2 = 2f_{XY}\sigma_u^2,$$

where σ_u^2 is the variance of additive genetic effects. Thus, the genetic correlation structure can be easily determined once the coefficients of coancestry are in hands.

The coefficient of coancestry for two individuals X and Y is defined as the **probability** that at a locus two randomly sampled alleles, one from either individual, are **identical by descent**, i.e., they stem from the same ancestor. Inbreeding generally increases the coefficient of coancestry. As a special case, the coefficient of coancestry of an inbred with itself is $f_{XX} = 1$. Some further typical values of f are as given in Table 12.

Table 12: Coefficient of coancestry, assuming unrelated parents (Bernardo, 2002).

Relationship	Coefficient of coancestry (f)	
	Parents non-inbred	Parents inbred
Parent - offspring	1/4	1/2
Full-sibs (offspring having the same parents)	1/4	1/2
Half-sibs (offspring having one common parent)	1/8	1/4

Pedigrees may become quite complex, and parents are often related. There are relatively straightforward ways to compute the coefficient of coancestry for general pedigrees. For example, if A and B are the parents of an individual X, while C and D are the parents of

individual Y, then the coefficient of coancestry between the two individuals X and Y is

$$f_{XY} = \frac{1}{4}\left(f_{AC} + f_{AD} + f_{BC} + f_{BD}\right) \quad ,$$

where f_{AC}, f_{AC}, f_{AC}, and f_{AC} are the coefficients of coancestry between two parents of different individuals. Here, we will simply assume that all coefficients of coancestry for a pedigree can be computed. For details see Bernardo (2002) or Falconer and Mackay (1996). It turns out that for computerized analysis, it is easier to compute a_{XY} instead of f_{XY}, as will be discussed next.

8.4 The numerator relationship matrix

The coefficients a_{XY} can be collected into a square matrix $\boldsymbol{A}$. This is known as the numerator relatonship matrix (Mrode, 1998). Assuming absence of dominance and epistasis, the variance-covariance matrix for genetic effects (breeding values) is given by

$$\boldsymbol{G} = \boldsymbol{A}\sigma_u^2 \quad .$$

The i-th diagonal element of matrix $\boldsymbol{A}$ is given by

$$a_{ii} = 1 + F_i \quad ,$$

where F_i is the inbreeding coefficient of the i-th genotype. If both parents (s and d) of genotype i are known, entries a_{ij} of $\boldsymbol{A}$ can be computed recursively (Henderson, 1976). Initially, individuals in the pedigree are coded 1 to n and ordered such that parents precede their progeny. The following rules are then employed to compute $\boldsymbol{A}$:

$a_{ji} = a_{ij} = 0.5(a_{js} + a_{jd}) \quad j = 1$ to $(i\text{-}1)$
$a_{ii} = 1 + 0.5(a_{sd})$

If only one parent (s) is known and assumed unrelated to the mate

$a_{ji} = a_{ij} = 0.5(a_{js}) \quad j = 1$ to $(i\text{-}1)$
$a_{ii} = 1$

If both parents are unknown and are assumed unrelated

$a_{ji} = a_{ij} = 0 \quad j = 1$ to $(i\text{-}1)$
$a_{ii} = 1$

Table 13: Pedigree for six genotypes (Mrode, 1998, p.26).

Offspring	Parent 1	Parent 2
3	1	2
4	1	unknown
5	4	3
6	5	2

The numerator relationship matrix for the pedigree in Table 10 is given by

	1	2	3	4	5	6
1	1	0	0.5	0.5	0.5	0.25
2	0	1	0.5	0	0.25	0.625
3	0.5	0.5	1	0.25	0.625	0.563
4	0.5	0	0.25	1	0.625	0.563
5	0.5	0.25	0.625	0.625	1.125	0.688
6	0.25	0.625	0.563	0.313	0.688	1.125

For instance:

$a_{11} = 1 + 0 = 1$
$a_{12} = 0.5(0 + 0) = 0 = a_{21}$
$a_{22} = 1 + 0 = 1$
$a_{13} = 0.5(a_{11} + a_{12}) = 0.5(1 + 0) = 0.5 = a_{31}$
$a_{23} = 0.5(a_{12} + a_{22}) = 0.5(0 + 1) = 0.5 = a_{32}$
$\vdots$
$a_{34} = 0.5(a_{13}) = 0.5(0.5) = 0.25 = a_{43}$
$\vdots$
$a_{66} = 1 + 0.52(a_{52}) = 1 + 0.5(0.25) = 1.125$

From the above calculation the inbreeding coefficient for genotype 6 is 0.125. The example shows that computation of the numerator relationship matrix $\boldsymbol{A}$ is straightforward, but somewhat tedious. Fortunately, there are computer programs available for this task, which merely require the pedigree in the form shown in Table 10. In the next section it is shown how $\boldsymbol{A}$ can be used to obtain BLUPs of genetic effects using pedigree information.

8.5 Pedigree-based BLUP

Example 16: For illustratioin we use a small example involving four barley genotypes. We will assume that these have been tested in a block experiment, and that the variance components are known. The coefficients of coancestry are given in Table 11.

Table 14: Coefficient of coancestry for four barley genotypes (Bernardo, 2002, p.222).

	Morex	Robust	Excel	Stander
Morex	1	1/2	7/16	11/32
Robust		1	27/32	43/64
Excel			1	91/128
Stander				1

With these coefficients of coancestry, the numerator relationship matrix is given by

$$\boldsymbol{A} = \begin{pmatrix} 2 & 1 & 7/8 & 11/16 \\ 1 & 2 & 27/16 & 43/32 \\ 7/8 & 27/16 & 2 & 91/64 \\ 11/16 & 43/32 & 91/64 & 2 \end{pmatrix}.$$

Now assume that the four genotypes have been tested in a randomized complete block design with three replicates. We analyse according to the linear model

$$y_{ij} = \mu + b_j + \tau_i + e_{ij} \quad ;$$

taking the genotype effects τ_i as random. The variance-covariance matrix is given by

$$\text{var}\begin{pmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \tau_4\end{pmatrix} = \boldsymbol{A}\sigma_u^2 = \begin{pmatrix} 2 & 1 & 7/8 & 11/16\\ 1 & 2 & 27/16 & 43/32\\ 7/8 & 27/16 & 2 & 91/64\\ 11/16 & 43/32 & 91/64 & 2\end{pmatrix}\sigma_u^2 = \boldsymbol{G} \qquad (2)$$

The observed means are assumed to be as follows:

Genotype	Mean
Morex	4.45
Robust	4.61
Excel	5.82
Stander	5.27
Overall mean	5.04

Now the means are correlated due to the genetic correlation of effects. The phenotypic covariance matrix, which we denote as ***P***, is the sum of the genetic covariance matrix ***G***, plus the error variance of a mean given by σ_e^2/r. It has the following form:

$$\text{var}\begin{pmatrix}\bar{y}_{1\bullet}\\ \bar{y}_{2\bullet}\\ \bar{y}_{3\bullet}\\ \bar{y}_{4\bullet}\end{pmatrix} = \boldsymbol{A}\sigma_u^2 + \boldsymbol{I}\frac{\sigma_e^2}{r} = \begin{pmatrix} 2 & 1 & 7/8 & 11/16\\ 1 & 2 & 27/16 & 43/32\\ 7/8 & 27/16 & 2 & 91/64\\ 11/16 & 43/32 & 91/64 & 2\end{pmatrix}\sigma_u^2 + \begin{pmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}\frac{\sigma_e^2}{r} = \boldsymbol{P} \quad .$$

Here, ***I*** denotes an identity matrix, which has 1s on the diagonal and 0s elsewhere. ***P*** is equal to the BLUPs of genetic effects can be shown to have the following form:

$$BLUP\begin{pmatrix}\tau_1\\ \tau_2\\ \tau_3\\ \tau_4\end{pmatrix} = \boldsymbol{G}\boldsymbol{P}^{-1}\begin{pmatrix}\bar{y}_{1\bullet}-\hat{\mu}\\ \bar{y}_{2\bullet}-\hat{\mu}\\ \bar{y}_{3\bullet}-\hat{\mu}\\ \bar{y}_{4\bullet}-\hat{\mu}\end{pmatrix} \quad ,$$

where $\hat{\mu}$ is a weighted overall mean of the form

$$\hat{\mu} = \left(\mathbf{1}'\boldsymbol{P}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{P}^{-1}\left(\bar{y}_{1\bullet}, \bar{y}_{2\bullet}, \bar{y}_{3\bullet}, \bar{y}_{4\bullet}\right)'$$

and thus $\bar{y}_{i\bullet} - \hat{\mu}$ is the **phenotypic deviation** of the *i*-th variety. The weighted mean $\hat{\mu}$ must be used here instead of a simple mean $\bar{y}_{\bullet\bullet}$ (as used with independent genotypes; see below) because $\boldsymbol{P}$ is not proportional to an identity matrix.

$\boldsymbol{P}^{-1}$ denotes the **inverse** of the matrix $\boldsymbol{P}$. An inverse is defined so that

$$\boldsymbol{P}^{-1}\boldsymbol{P} = \boldsymbol{I} \ .$$

Multiplication by an inverse corresponds to a division in scalar algebra. Note the analogy with the BLUP formula (1) in Section 3.3.2, which we repeat here:

$$BLUP(\tau_i) = h^2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \ .$$

While in section 3.3.2 the phenotypic deviations $(\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet})$ were multiplied by h^2, the ratio of genetic and phenotypic variance, we here multiply them by $\boldsymbol{G}\boldsymbol{P}^{-1}$, which, loosely speaking, is a generalized ratio of genetic and phenotypic variances. In fact, the two equations become identical of $\boldsymbol{G}$ is replaced by a diagonal matrix, i.e., if we assume that the genotypes are uncorrelated. This assumption was, in fact, made in section 3.3.2. In order to study the effect of genetic correlation, we will compute BLUP under both assumtions here.

If it is assumed that $\sigma_u^2 = 0.2$ and $\sigma_e^2 = 1.0$ we find

$$\boldsymbol{G} = \begin{pmatrix} 0.4 & 0.2 & 0.175 & 0.1375 \\ 0.2 & 0.4 & 0.3375 & 0.26875 \\ 0.175 & 0.3375 & 0.4 & 0.284375 \\ 0.1375 & 0.26875 & 0.284375 & 0.4 \end{pmatrix} \text{ and}$$

$$\boldsymbol{P} = \begin{pmatrix} 1.4 & 0.2 & 0.175 & 0.1375 \\ 0.2 & 1.4 & 0.3375 & 0.26875 \\ 0.175 & 0.3375 & 1.4 & 0.284375 \\ 0.1375 & 0.26875 & 0.284375 & 1.4 \end{pmatrix} \text{ and}$$

$$\boldsymbol{G}\boldsymbol{P}^{-1} = \begin{pmatrix} 0.262 & 0.082 & 0.064 & 0.044 \\ 0.082 & 0.215 & 0.157 & 0.111 \\ 0.064 & 0.157 & 0.215 & 0.123 \\ 0.044 & 0.111 & 0.123 & 0.235 \end{pmatrix}$$

$$BLUP\begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \boldsymbol{G}\boldsymbol{P}^{-1}\begin{pmatrix} \bar{y}_{1\bullet} - \hat{\mu} \\ \bar{y}_{2\bullet} - \hat{\mu} \\ \bar{y}_{3\bullet} - \hat{\mu} \\ \bar{y}_{4\bullet} - \hat{\mu} \end{pmatrix} = \begin{pmatrix} 0.262 & 0.082 & 0.064 & 0.044 \\ 0.082 & 0.215 & 0.157 & 0.111 \\ 0.064 & 0.157 & 0.215 & 0.123 \\ 0.044 & 0.111 & 0.123 & 0.235 \end{pmatrix}\begin{pmatrix} -0.5613 \\ -0.4013 \\ +0.8087 \\ +0.2587 \end{pmatrix} = \begin{pmatrix} -0.1169 \\ +0.0231 \\ +0.1072 \\ +0.0913 \end{pmatrix} \ .$$

For comparison, we ignore genetic correlation, in which case both $\boldsymbol{G}$ and $\boldsymbol{P}$ become diagonal.

$$\boldsymbol{G} = \begin{pmatrix} 0.4 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.4 \end{pmatrix} \text{ and } \boldsymbol{P} = \begin{pmatrix} 1.4 & 0 & 0 & 0 \\ 0 & 1.4 & 0 & 0 \\ 0 & 0 & 1.4 & 0 \\ 0 & 0 & 0 & 1.4 \end{pmatrix}, \text{ so that}$$

$$\boldsymbol{GP}^{-1} = \begin{pmatrix} 0.286 & 0 & 0 & 0 \\ 0 & 0.286 & 0 & 0 \\ 0 & 0 & 0.286 & 0 \\ 0 & 0 & 0 & 0.286 \end{pmatrix} \text{ and}$$

$$BLUP\begin{pmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \boldsymbol{GP}^{-1}\begin{pmatrix} \bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet} \\ \bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet} \\ \bar{y}_{3\bullet} - \bar{y}_{\bullet\bullet} \\ \bar{y}_{4\bullet} - \bar{y}_{\bullet\bullet} \end{pmatrix} = \begin{pmatrix} 0.286 & 0 & 0 & 0 \\ 0 & 0.286 & 0 & 0 \\ 0 & 0 & 0.286 & 0 \\ 0 & 0 & 0 & 0.286 \end{pmatrix}\begin{pmatrix} -0.5875 \\ -0.4275 \\ +0.7825 \\ +0.2325 \end{pmatrix} = \begin{pmatrix} -0.1679 \\ -0.1221 \\ +0.2236 \\ +0.0664 \end{pmatrix}.$$

Note that this is equivalent to using $BLUP(\tau_i) = h^2(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$ with $h^2 = 0.4/1.4$. The rank order of both sets of BLUPs are identical, but estimates differ notably. For example, the sign for the BLUP of τ_2 (Morex) has changed. Also, the BLUPs of effects τ_3 and τ_4 of the last two genotypes (Excel and Stander) are much closer when genetic correlation is accounted for.

The simple example has shown that genetic correlation matters and has a notable effect on BLUP. Thus, genetic correlation should not be ignored. The good news is that accounting for genetic correlation will tend to increase accuracy of estimates, because information from relatives is exploited. Thus, a small number of replicates per genotype can be partly counterbalanced by lending information from relatives.

The BLUP equation we have used here applies only to data from a randomized complete block design. With incomplete blocking, BLUP equations will become more complicated due to the need to adjust for block effects. We will not give these more general BLUP equations here (but see Appendix A to get a feel), because they are heavily laden with matrix algebra (Bernardo, 2002) and do not provide new insights at this point. The centrepiece of these equations remains the numerator relationship matrix, and this fact is important for the user of a statistical package. In using such a package the crucial step is to feed the numerator relationship matrix into the program. In so doing it is important to note that the genetic covariance will be of the form as given in eq. (2), no matter how complex the blocking structure of the design.

Example 17 (C. Flachenecker, Universität Hohenheim, pers. comm., **CxD03ewe.dat**): Two European flint lines (maize), denoted here as C and D, were used to generate F_2 populations from cross C × D. The F_2syn_3 generation was obtained by chain crossing of F_2-plants. F_2syn_3 plants were then used to generate full-sib (FS) families. The resulting 142 FS families were tested in an α-design with three replicates. The data were analysed according to the model for an α-design given in 3.2.8:

$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + e_{ijh}$.

The numerator relationship matrix $\boldsymbol{A}$ was generated from the pedigree using the INBREED procedure of SAS (**CD7ver.xls**). The upper right 4 × 4 block of this matrix is as follows:

$$\boldsymbol{A} = \begin{pmatrix} 1.086 & 0.610 & 0.398 & 0.418 & . & . \\ 0.610 & 1.235 & 0.477 & 0.505 & . & . \\ 0.398 & 0.477 & 1.080 & 0.534 & . & . \\ 0.418 & 0.505 & 0.534 & 1.108 & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix} .$$

The genetic covariance matrix for the genotypic effects was assumed to be

$$\text{var}\begin{pmatrix} \tau_1 \\ \tau_2 \\ . \\ . \end{pmatrix} = \boldsymbol{A}\sigma_u^2 = \boldsymbol{G} \quad .$$

Block effects were taken as random for recovery of inter-block information. The resulting variance component estimates were:

Error	$\hat{\sigma}_e^2 = 30.19$	Akaike Information Criterion (AIC) = 2870.4
Blocks	$\hat{\sigma}_b^2 = 6.77$	
Genetic effects	$\hat{\sigma}_u^2 = 49.32$	

For comparison, we analysed the data assuming that genetic effects were independent. The variance component estimates were:

Error	$\hat{\sigma}_e^2 = 29.81$	AIC = 2881.4
Blocks	$\hat{\sigma}_b^2 = 7.07$	
Genetic effects	$\hat{\sigma}_\tau^2 = 36.12$	

The Akaike Information Criterion (AIC) is a measure for the goodness-of-fit of a model. The smaller the AIC the better the fit. According to AIC, the model exploiting the pedigree fits better than the model ignoring this information. The rank correlation of BLUPs computed by both models is 0.98. Thus, rank orders are similar, but not identical, and selection decisions would therefore not be identical.

8.6 BLUP under selection

Estimation of genetic effects by BLUP assumes that genotypes in a pedigree can be regarded as a random sample. The laws of Mendelian segregation justify this assumption. It is important to note, however, that in breeding applications, genotypes are subject to selection in each generation. Thus, not all genotypes in the pedigree that would have been observed without selection, will be observed in practice.

Fortunately, BLUP remains valid under fairly general conditions, even under selection. The most general condition for the validity of BLUP is that all information that was employed in selection decisions is included in the analysis (Piepho and Möhring, 2006). This will necessitate the joint analysis of data from several years (see lecture by F.A. van Eeuwijk). Also, selection will usually be based on several traits. It is therefore adequate to perform a multivariate analysis, at least including the most important traits. Multivariate BLUP is closely related to the use of index selection. This type of analysis is quite common in animal breeding, but not yet quite as common in plant breeding.

8.7 Multivariate BLUP

It is often preferable to jointly analyse several traits. This allows an assessment of genetic and phenotypic correlation among traits, which may in turn be useful to study the response to indirect selection. Also, BLUPs under a multivariate model are typically more efficient than univariate BLUPs because correlation is exploited to extract information on a target trait from correlated traits. Finally, multivariate BLUP is preferable under selection (see. 3.3.6).

Example 8 (cont'd): A yield trial with 64 oat genotypes was laid out as a 8 x 8 lattice with 3 replicates (H. F. Utz, pers. comm.; **rowcol_fromutz.dat**). The data may be analysed using the model

$$y_{ijh} = \mu + \gamma_j + b_{jh} + \tau_i + e_{ijh}$$

which was described in section 3.2.8. We here take block effects (b_{jh}) as fixed, but take genotype effects (τ_i) as random. The dataset has three different traits: yield (**`yield`**), thousand kernel weight (**`TKW`**), and plant height (**`height`**). One may study the correlation of these traits. Under the above mixed model, this may be done separately for each random effect. It is also helpful to partition genetic and environmental sources of covariance, e.g., to assess the reponse to indirect selection and to compute optimal selection indices. Specifically, the genetic correlation or covariance can be assessed based on the genetic effects for the different traits. The variances and covariances may be expressed as

$$\text{var}\begin{pmatrix}\tau_{i1}\\ \tau_{i2}\\ \tau_{i3}\end{pmatrix}=\begin{pmatrix}\sigma^2_{\tau 1} & \sigma_{\tau 12} & \sigma_{\tau 13}\\ \sigma_{\tau 12} & \sigma^2_{\tau 2} & \sigma_{\tau 23}\\ \sigma_{\tau 13} & \sigma_{\tau 23} & \sigma^2_{\tau 3}\end{pmatrix}$$

where $\sigma^2_{\tau t}$ is the genetic variance of the t-th trait and $\sigma_{\tau t t'}$ is the genetic covariance among traits t and t'. The genetic correlation among traits 1 and 2 is defined as

$$\rho_{\tau 12}=\frac{\sigma_{\tau 12}}{\sqrt{\sigma^2_{\tau 1}\sigma^2_{\tau 2}}}$$

Deleting three records with missing data for yield, we find the following estimates:

$$\hat{\text{var}}\begin{pmatrix}\tau_{i1}(yield)\\ \tau_{i2}(TKW)\\ \tau_{i3}(height)\end{pmatrix}=\begin{pmatrix}\hat{\sigma}^2_{\tau 1} & \hat{\sigma}_{\tau 12} & \hat{\sigma}_{\tau 13}\\ \hat{\sigma}_{\tau 12} & \hat{\sigma}^2_{\tau 2} & \hat{\sigma}_{\tau 23}\\ \hat{\sigma}_{\tau 13} & \hat{\sigma}_{\tau 23} & \hat{\sigma}^2_{\tau 3}\end{pmatrix}=\begin{pmatrix}29623 & 2.083 & 425.1\\ 2.083 & 4.007 & 3.771\\ 425.1 & 3.771 & 22.33\end{pmatrix}$$

From this the genetic correlation is computed as follows:

$$\hat{\rho}_{\tau 12}=\frac{\hat{\sigma}_{\tau 12}}{\sqrt{\hat{\sigma}^2_{\tau 1}\hat{\sigma}^2_{\tau 2}}}=\frac{2.083}{\sqrt{29623*4.007}}=0.006046$$

The full genetic correlation matrix is

$$\hat{corr}\begin{pmatrix}\tau_{i1}(yield)\\ \tau_{i2}(TKW)\\ \tau_{i3}(height)\end{pmatrix}=\begin{pmatrix}1 & \hat{\rho}_{\tau 12} & \hat{\rho}_{\tau 13}\\ \hat{\rho}_{\tau 21} & 1 & \hat{\rho}_{\tau 23}\\ \hat{\rho}_{\tau 31} & \hat{\rho}_{\tau 32} & 1\end{pmatrix}=\begin{pmatrix}1 & 0.006 & 0.523\\ 0.006 & 1 & 0.399\\ 0.523 & 0.399 & 1\end{pmatrix}$$

The environmental covariance based on error effects can be defined as

$$\text{var}\begin{pmatrix}e_{ij1}\\ e_{ij2}\\ e_{ij3}\end{pmatrix}=\begin{pmatrix}\sigma^2_{e1} & \sigma_{e12} & \sigma_{e13}\\ \sigma_{e12} & \sigma^2_{e2} & \sigma_{e23}\\ \sigma_{e13} & \sigma_{e23} & \sigma^2_{e3}\end{pmatrix}$$

The REML estimate is

$$\hat{\text{var}}\begin{pmatrix}e_{ij1}\\ e_{ij2}\\ e_{ij3}\end{pmatrix}=\begin{pmatrix}19316 & 70.93 & 107.2\\ 70.93 & 5.937 & -1.199\\ 107.2 & -1.199 & 8.855\end{pmatrix}$$

A phenotypic variance-covariance matrix on a plot basis is obtained by adding the two variance-covariance matrices:

$$\hat{\text{var}}\begin{pmatrix}\tau_{i1}+e_{ij1}\\ \tau_{i2}+e_{ij2}\\ \tau_{i3}+e_{ij3}\end{pmatrix}=\begin{pmatrix}29623 & 2.083 & 425.1\\ 2.083 & 4.007 & 3.771\\ 425.1 & 3.771 & 22.33\end{pmatrix}+\begin{pmatrix}19316 & 70.93 & 107.2\\ 70.93 & 5.937 & -1.199\\ 107.2 & -1.199 & 8.855\end{pmatrix}=\begin{bmatrix}48939 & 73.013 & 532.3\\ 73.013 & 9.944 & 2.572\\ 532.3 & 2.572 & 31.155\end{bmatrix}$$

From this, the phenotypic correlation on a plot basis is

$$\hat{corr}\begin{pmatrix}\tau_{i1}+e_{ij1}\\ \tau_{i2}+e_{ij2}\\ \tau_{i3}+e_{ij3}\end{pmatrix}=\begin{pmatrix}1 & 0.105 & 0.431\\ 0.105 & 1 & 0.146\\ 0.431 & 0.146 & 1\end{pmatrix}$$

We may compute the BLUPs of genetic effects for the different traits under the multivariate mixed model. These BLUPs exploit genetic correlation in order to obtain more accurate

estimates. The resulting estimates differ from the univariate BLUPs, which are obtained by a trait-by-trait analysis.

	Multivariate BLUP			Univariate BLUP		
variable	yield	TKW	height	yield	TKW	height
Genotype						
1	65.64	0.47	6.54	39.11	-0.31	6.76
2	-252.22	1.79	-4.31	-236.03	1.82	-4.70
3	-131.77	1.98	-2.85	-104.28	2.23	-3.56
4	124.46	-0.13	3.33	115.21	-0.30	3.38
5	31.02	3.13	6.09	33.61	2.68	5.48
6	226.25	3.19	6.35	241.30	3.22	5.53
7	9.53	1.71	2.88	15.31	1.60	2.52
8	60.86	-1.78	2.21	37.85	-2.15	2.71
9	-132.66	-3.70	-5.29	-152.78	-3.74	-4.34
10	-53.78	-0.87	4.18	-90.30	-1.72	4.89
11	-33.78	-1.33	0.53	-47.89	-1.54	1.01
12	29.56	-1.24	-4.45	45.82	-0.60	-4.54
13	-79.49	0.67	-0.66	-75.39	0.67	-0.73
14	-80.51	-1.82	-7.14	-69.02	-1.22	-6.88
15	-65.72	0.71	-7.00	-23.11	1.69	-7.68
16	-73.46	0.38	-4.88	-47.04	0.97	-5.24
17	-38.75	-0.07	0.80	-46.58	-0.27	0.95
18	196.73	2.29	6.71	196.24	2.13	6.34
19	21.06	-1.24	-4.31	36.99	-0.68	-4.44
20	72.34	0.33	1.25	78.10	0.46	1.03
21	15.46	-2.20	-3.33	12.83	-1.90	-2.97
22	111.60	-2.37	-5.76	125.93	-1.56	-5.69
23	-58.77	-1.11	-4.91	-44.29	-0.59	-4.92
24	-3.49	-0.79	-2.18	4.07	-0.47	-2.17
25	-58.52	0.13	-3.06	-52.08	0.31	-3.14
26	301.84	0.06	8.76	275.45	-0.49	8.87
27	-55.44	3.88	4.31	-42.47	3.56	3.53
28	200.05	2.26	6.46	201.70	2.09	5.93
29	161.51	0.84	4.53	156.75	0.64	4.27
30	-53.05	-0.98	2.51	-83.04	-1.64	3.13
31	140.65	-1.46	2.39	125.51	-1.59	2.65
32	84.55	-2.03	-4.38	98.24	-1.23	-4.21
33	107.90	0.27	1.05	110.54	0.38	0.87
34	142.58	0.40	4.54	127.95	0.09	4.59
35	18.11	-0.52	-0.71	18.63	-0.40	-0.63
36	197.82	-0.77	4.55	182.26	-0.95	4.70
37	201.24	-0.38	3.31	194.08	-0.39	3.28
38	180.91	-1.39	1.70	169.65	-1.30	2.00
39	158.68	-1.61	-4.14	179.90	-0.72	-4.40
40	88.72	0.39	3.48	80.62	0.22	3.55
41	77.93	0.96	6.38	56.58	0.35	6.56
42	82.73	-0.33	5.30	50.81	-1.00	5.85
43	-299.52	-0.28	-4.72	-300.66	-0.43	-4.46
44	39.41	-1.42	-2.74	40.12	-1.13	-2.55
45	117.44	-0.88	-1.21	120.68	-0.54	-1.17
46	-93.49	-1.95	-5.06	-97.47	-1.78	-4.64
47	76.51	-1.19	-3.39	98.82	-0.45	-3.68
48	-55.87	0.14	2.35	-71.44	-0.31	2.64
49	-316.97	-0.80	-3.62	-332.60	-1.26	-3.07
50	-119.85	1.14	1.12	-127.03	0.66	1.06
51	-140.39	3.32	2.55	-130.06	2.96	2.03
52	-142.49	2.89	1.40	-132.95	2.63	0.93
53	-204.27	1.14	-0.84	-203.00	0.84	-0.97
54	-69.80	1.75	5.92	-88.69	0.89	5.99
55	193.31	1.94	4.08	207.95	2.07	3.43
56	-138.34	-1.61	-4.18	-143.00	-1.60	-3.77
57	188.01	1.76	4.71	192.79	1.75	4.35
58	-94.51	-1.74	-1.80	-109.91	-2.00	-1.31
59	-86.89	-0.78	-7.93	-54.36	0.07	-8.27
60	43.28	-2.94	-6.83	55.26	-2.16	-6.61
61	156.75	-0.16	2.19	156.44	-0.08	2.16
62	-536.21	-0.16	-5.57	-553.03	-0.88	-4.69
63	-180.34	-0.46	-1.58	-186.10	-0.63	-1.23
64	-274.10	2.60	-5.62	-238.51	3.03	-6.27

Final remark: A review on the use of BLUP for random genetic effects in plant breeding is found in Piepho et al. 2008a).

8.8 GCA and SCA in a diallel

Example 19: Consider a small diallel with five parents (simulated data). Means for the crosses are given in the following table.

Female parent	Male parent 1	2	3	4	5
1					
2	62.53				
3	55.92	64.54			
4	61.97	75.69	77.15		
5	60.02	75.15	71.34	70.89	

The dataset (**diallel.dat**) has two replicates, and the design is completely randomized (no blocks).

The model for the genetic effects may be stated as

$$g_{ij} = (g.c.a.)_i + (g.c.a.)_j + (s.c.a.)_{ij},$$

Where g_{ij} is the genetic effect of the cross of parents i and j, $(g.c.a.)_i$ is the g.c.a. effect of the i-th parent, and $(s.c.a.)_{ij}$ is the s.c.a. effect of the cross between parents i and j.

A problem in setting up a mixed model for g_{ij} is that the g.c.a. effect appears twice in the model. Specifically, the same g.c.a. effect may appear as the first g.c.a. effect $(g.c.a.)_i$ for some crosses and as the second g.c.a. effect $(g.c.a.)_j$ in some other crosses. We could set up a dataset with two factors **`parent1`** and **`parent2`**, which code for the two parents and then formulate a simple two-way model

```
parent1 + parent2 + parent1.parent2
```

,

but this is not appropriate. The same effect will need to be modelled by both **`parent1`** and **`parent2`**. For a diallel with five parents we have:

Cross	Parent1	Parent2
1x2	1	2
1x3	1	3
1x4	1	4
1x5	1	5
2x3	2	3
2x4	2	4
2x5	2	5
3x4	3	4
3x5	3	5
4x5	4	5

For example, parent 2 appears as **parent2** in the cross 1x2 and as **parent1** in the cross 2x3. Clearly, the effect to be modelled is the same in both cases: $(g.c.a.)_2$. But a simple two-way model will assign different effects to the factors **parent1** and **parent2**. In case of a diallel with five parents, 10 effects will be modelled instead of three.

One way to tackle the problem is to define dummy variables, one for each g.c.a. effect. For three parents, the dummy coding of dummy variables **gca1**, **gca2** and **gca3** is as follows:

Cross	**parent1**	**parent2**	**gca1** x_1	**gca2** x_2	**gca3** x_3	**gca4** x_4	**gca5** x_5
1x2	1	2	1	1	0	0	0
1x3	1	3	1	0	1	0	0
1x4	1	4	1	0	0	1	0
1x5	1	5	1	0	0	0	1
2x3	2	3	0	1	1	0	0
2x4	2	4	0	1	0	1	0
2x5	2	5	0	1	0	0	1
3x4	3	4	0	0	1	1	0
3x5	3	5	0	0	1	0	1
4x5	4	5	0	0	0	1	1

Random coefficients can then be fitted for the dummies according to the model

$$g_{ij} = \sum_{i=1}^{p} x_i (g.c.a.)_i + (s.c.a.)_{ij},$$

where x_i is the dummy variable for the i-th parent and p is the number of parents, and a constant variance be inforced for all random coefficients $(g.c.a.)_i$. In GenStat this model is very easily fitted by defining the dummies with numerals in brackets: **gca[1]**, **gca[2]**, …, and then fitting a single random effect **gca**.

Example 19 (continued): For the simulated data we find the following variance component estimates:

Effect	Estimate
g.c.a.	28.6644
s.c.a.	16.1710
Residual	2.1184

Remark: Another design that also allows estimation of g.c.a. and s.c.a. variances, is a **factorial design**. In this design, two heterotic pools A and B are considered. Several crosses are performed between parents of the one pool and parents of the other pool. Now a separate variance for g.c.a. needs to be defined for each pool, and the appropriate linear model is just a standard two-way ANOVA model with main effects A and B corresponding to g.c.a. effects for the two pools A and B, and the interaction A.B corresponding to s.c.a. effects.

***8.9 The general BLUP equation and the mixed model equations**

Consider the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

with $\boldsymbol{e} \sim \text{MNV}(\boldsymbol{0},\boldsymbol{R})$, $\boldsymbol{u} \sim \text{MNV}(\boldsymbol{0},\boldsymbol{G})$, and $\boldsymbol{y} \sim \text{MNV}(\boldsymbol{X\beta}, \boldsymbol{V})$, where $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$. The estimation of the random effect $\boldsymbol{u}$ by **Best Linear Unbiased Prediction (BLUP)** is obtained by

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{G}}\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

Both BLUE and BLUP may be computed, and usually are due to better computational efficiency, by solving the so-called **Mixed Model Equations (MME)**, given by (Searle et al., 1992)

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

where $\boldsymbol{G}^{-1}$ and $\boldsymbol{R}^{-1}$ are the inverses of $\boldsymbol{G}$ and $\boldsymbol{R}$, respectively. It is useful to observe that when the variances in $\boldsymbol{G}$ become very large then $\boldsymbol{G}^{-1}$ becomes tiny, and the mixed model equations become

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

showing that for large variance random effects essentially turn into fixed effects. If, furthermore, residual errors are independent with homogeneous variance, i.e., $\boldsymbol{R}^{-1} = \sigma^{-2}\boldsymbol{I}$, the MME turn into the **ordinary least squares equations**

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

Example 3 (continued): When the block variance is very large, analysis with random blocks is essentially the same as analysis with fixed blocks, where only the intra-block information is used and no inter-block information is recovered.

9 Series of experiments

9.1 Basic modelling

Example 20: The German variety testing office (Bundessortenamt) tests new cultivars at a number of locations each year in order to assess their value for cultivation and use (VCU). Results on mean performance are published yearly so that growers and extension services can make informed cultivar choices. Trait means are based on the assumption that locations are a random sample of locations in the target population of environments. Analysis can be performed based on a mixed model with random effects for locations and fixed effects for cultivars. This will be illustrated using data from 18 trials with 14 oats cultivars (**oats.dat**). All trials were laid out in randomized complete blocks with four replicates.

Objective: Computation of means over locations, using a suitable variance-covariance structure for weighting according to different sources of variation.

Modelling: A simple ANOVA model can be derived from the linear model for a single trial given by

$$y_{ik} = \mu + \tau_i + b_k + e_{ik}$$

where

y_{ik} = yield of i-th cultivar in k-th block
μ = general mean
τ_i = effect of i-th cultivar
b_k = effect of k-th block
e_{ik} = error associated with y_{ik}

Extending the model to a series of trials at different locations, we simple add an index for locations (j):

$$y_{ijk} = \mu_j + \tau_{ij} + b_{jk} + e_{ijk}$$

where

y_{ijk} = yield of i-th cultivar in k-th block in j-th location
μ_j = general mean in j-th location
τ_{ij} = effect of i-th cultivar in j-th location
b_{jk} = effect of k-th block in j-th location
e_{ijk} = error associated with y_{ijk}

The effects μ_j and τ_{ij} are confounded. They represent effects of the factors cultivar and location. These effects can be replaced by a two-factorial model:

$$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where

μ = general mean

α_i = main effect of i-th cultivar
β_j = main effect of j-th location
$(\alpha\beta)_{ij}$ = interaction of i-th cultivar and j-th location

Plugging this model into the overall model in place of effects μ_j and τ_{ij} yields:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_{jk} + e_{ijk} \quad .$$

Then the location factor is random, then conventionally all effects associated with this factor are regarded as random: the location main effect, the cultivar × location interaction, and the block effect that is nested within locations. Thus, the distributional assumptions are:

$$\beta_j \sim N\left(0, \sigma_\beta^2\right)$$
$$(\alpha\beta)_{ij} \sim N\left(0, \sigma_{\alpha\beta}^2\right)$$
$$b_{jk} \sim N\left(0, \sigma_b^2\right)$$
$$e_{ijk} \sim N\left(0, \sigma^2\right)$$

Heterogeneity of variance: This model implies homogeneity of variance between locations. While this may seem a strong assumption, it can be justified by randomization theory (Calinski et al., 2005). Some gain in efficiency is possible, however, by accounting for heterogeneity of variance. In practice, differences in precision between trials are often quite pronounced.

In order to check for heterogeneity of variance, a likelihood ratio test may be performed, which compares the simple model with an extended model assuming

$$e_{ijk} \sim N\left(0, \sigma_j^2\right) \quad .$$

For each model we compute the restricted log likelihood (log L_R) and compute

$$T = -2\ [\log L_R(\text{homogeneous}) - \log L_R(\text{heterogeneous})] \quad .$$

Under the null hypothesis of homogeneous variance the test statistic T has a χ^2-distribution with (J–1) degrees of freedom, where J is the number of locations.

Example 20 (continued): For the oats data we find:

Model	$-2 \log L_R$
Homogeneous variances	3808.8
Heterogeneous variances	3644.5
	$T = 164.3$

The LR-statistic is significant on (J – 1) = 17 degrees of freedom (critical value at $\alpha = 5\%$: 27.59), showing that there is heterogeneity of variance.

We analyse the data based on a model with heterogeneous error variances. All effects except α_i are taken as random, because the location factor is random. The analysis here is in a single stage and will later be contrasted to a two-stage analysis, where in a first step trial means per cultivar are cumputed, which are then summarized across trials using a suitable model. There are marked differences among error variances. By comparison, the data were also analysed assuming homogeneous error variances. Differences are minor, the s.e.d. being slightly bigger than with heterogeneous variances.

Fixed locations and blocks: An alternative analysis takes location and block effects (β_j and b_{jk}) as fixed. This is suitable when only contrasts among cultivars are of interest, which is usually the case. We may take these effects fixed, because they are not crossed in any way with the effect of interest: cultivars. Thus, we may not take interactions and errors [$(\alpha\beta)_{ij}$ and errors e_{ijk}] fixed, as they share an index (i) with cultivar effects. The situation is similar to analysis of an incomplete block design, where we may take blocks fixed or random. With random blocks, we can exploit inter-block information, if any. Similarly, analysis of a series of trials with fixed location effects means that we disregard inter-location information. This information is generally small, because the variance of location effects is large. In the present case, there is no inter-location information, because the design is orthogonal.

2-stage-analysis: Single-stage analysis may be computationally demanding. It is then interesting to proceed in two stages instead. In the first step, means are computed per cultivar and location. These means are then subjected to analysis by a two-way model. In the present case the model is:

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ij}$$

where

y_{ij} = mean of i-th cultivar in j-th location
e_{ij} = error of mean y_{ij}

In the first step we may compute the variance of a mean y_{ij} and use these to model errors e_{ij} in stage two. In stage two block effects are then implicitly accounted for as random effects, because block means per location are confounded with location main effects.

Example 20 (continued): Numerical differences with single-stage analysis are minor (Table 15). For this reason, we subsequently use two-stage analysis to explore extension of the model for cultivar-location effects.

Table 15: Means and standard errors based on different mixed models for oats data.

Analysis	1-stage				2-stage	
Random effects	all except α_i		only $(\alpha\beta)_{ij}$ and e_{ijk}		all except α_i	only $(\alpha\beta)_{ij}$ and e_{ijk}
Error variance	Homog.	Heterog.	Homog.	Heterog.	Heterog.	Heterog.
1	53.65	53.70	53.65	53.71	53.70	53.70
2	51.07	51.05	51.07	51.05	51.06	51.06
3	55.67	55.69	55.67	55.69	55.72	55.72
4	56.24	56.54	56.24	56.54	56.53	56.53
5	55.98	55.86	55.98	55.86	55.85	55.85
6	53.61	53.58	53.61	53.58	53.59	53.59
7	53.81	53.82	53.81	53.83	53.82	53.82
8	52.37	52.19	52.37	52.19	52.16	52.16
9	58.47	58.58	58.47	58.58	58.59	58.59
10	56.58	56.54	56.58	56.54	56.56	56.56
11	53.86	53.90	53.86	53.90	53.88	53.88
12	57.24	56.96	57.24	56.97	56.97	56.97
13	54.72	54.84	54.72	54.84	54.83	54.83
14	53.77	53.78	53.77	53.78	53.79	53.79
s.e.m.[$]	2.27	2.27	0.68	0.67	2.27	0.68
s.e.d.[§]	0.96	0.95	0.96	0.95	0.96	0.96

$ s.e.d. = standard error of a mean
§ s.e.d. = standard error of a difference

Table 16: Variance component estimates for oats data.

Effect	Variance component	Estimate
Main effect location	σ^2_β	82.34
Cultivar × location interaction	$\sigma^2_{\alpha\beta}$	6.07
Block within location	σ^2_b	7.74
Error location 1	σ^2_1	3.29
Error location 2	σ^2_2	5.39
Error location 3	σ^2_3	11.04
Error location 4	σ^2_4	6.00
Error location 5	σ^2_5	3.55
Error location 6	σ^2_6	9.76
Error location 7	σ^2_7	5.95
Error location 8	σ^2_8	4.73

Error location 9	σ^2_9	9.71
Error location 10	σ^2_{10}	15.21
Error location 11	σ^2_{11}	6.35
Error location 12	σ^2_{12}	5.36
Error location 13	σ^2_{13}	7.20
Error location 14	σ^2_{14}	7.95
Error location 15	σ^2_{15}	44.07
Error location 16	σ^2_{16}	4.90
Error location 17	σ^2_{17}	7.64
Error location 18	σ^2_{18}	11.11

9.2 Modelling cultivar × location interaction

We focus here on the cultivar × location part of the model, which is

$\eta_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$,

where η_{ij} is the conditional expectation of the *i*-th cultivar in the *j*-th location. So far, we have assumed

$$\beta_j \sim N\left(0, \sigma^2_\beta\right)$$
$$(\alpha\beta)_{ij} \sim N\left(0, \sigma^2_{\alpha\beta}\right)$$

This implies a relatively simple variance-covariance structure for η_{ij}, known as "**compound symmetry**":

$$\text{var}\left(\eta_{ij}\right) = \sigma^2_\beta + \sigma^2_{\alpha\beta}$$
$$\text{cov}\left(\eta_{ij}, \eta_{i'j}\right) = \sigma^2_\beta$$

Note that in repeated measures terminology, locations can be regarded as subjects, while cultivars correspond to time points. We exemplify the matrix representation only for 4 cultivars instead of 18 to save space:

$$\text{var}\begin{pmatrix}\eta_{1j}\\ \eta_{2j}\\ \eta_{3j}\\ \eta_{4j}\end{pmatrix} = \begin{pmatrix}\sigma^2_\beta + \sigma^2_{\alpha\beta} & \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta\\ \sigma^2_\beta & \sigma^2_\beta + \sigma^2_{\alpha\beta} & \sigma^2_\beta & \sigma^2_\beta\\ \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta + \sigma^2_{\alpha\beta} & \sigma^2_\beta\\ \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta & \sigma^2_\beta + \sigma^2_{\alpha\beta}\end{pmatrix}$$

In particular, each cultivar has the same variance, and each pair of cultivars has the same covariance. In practice, one often finds “stability” differences, meaning that variances are heterogeneous. This can be modelled as follows:

$$\beta_j \sim N\left(0, \sigma_\beta^2\right)$$
$$(\alpha\beta)_{ij} \sim N\left(0, \sigma_{\alpha\beta(i)}^2\right)$$

and

$$\text{var}\left(\eta_{ij}\right) = \sigma_\beta^2 + \sigma_{\alpha\beta(i)}^2$$
$$\text{cov}\left(\eta_{ij}, \eta_{i'j}\right) = \sigma_\beta^2$$

The variance $\sigma_{\alpha\beta(i)}^2$ is Shukla's **stability variance**. In matrix form we have

$$\text{var}\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \\ \eta_{4j} \end{pmatrix} = \begin{pmatrix} \sigma_\beta^2 + \sigma_{\alpha\beta(1)}^2 & \sigma_\beta^2 & \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma_\beta^2 + \sigma_{\alpha\beta(2)}^2 & \sigma_\beta^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma_\beta^2 & \sigma_\beta^2 + \sigma_{\alpha\beta(3)}^2 & \sigma_\beta^2 \\ \sigma_\beta^2 & \sigma_\beta^2 & \sigma_\beta^2 & \sigma_\beta^2 + \sigma_{\alpha\beta(4)}^2 \end{pmatrix}$$

The model is still restrictive in that a constant covariance is assumed. The most flexible model is the **unstructured model**, where each pair of cultivars has its own covariance. The model can be written

$\eta_{ij} = \mu + \alpha_i + \gamma_{ij}$

where

γ_{ij} = deviation for j-th location of η_{ij} from expected value of the i-th cultivar ($\mu + \alpha_i$).

$$\text{var}\left(\eta_{ij}\right) = \text{var}\left(\gamma_{ij}\right) = \sigma_{\gamma(i)}^2 = \sigma_{\gamma(ii)}$$
$$\text{cov}\left(\eta_{ij}, \eta_{i'j}\right) = \text{cov}\left(\gamma_{ij}, \gamma_{i'j}\right) = \sigma_{\gamma(ii')}$$

or in matrix form:

$$\text{var}\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \\ \eta_{4j} \end{pmatrix} = \begin{pmatrix} \sigma_{\gamma(11)} & \sigma_{\gamma(12)} & \sigma_{\gamma(13)} & \sigma_{\gamma(14)} \\ \sigma_{\gamma(21)} & \sigma_{\gamma(22)} & \sigma_{\gamma(23)} & \sigma_{\gamma(24)} \\ \sigma_{\gamma(31)} & \sigma_{\gamma(32)} & \sigma_{\gamma(33)} & \sigma_{\gamma(34)} \\ \sigma_{\gamma(41)} & \sigma_{\gamma(42)} & \sigma_{\gamma(43)} & \sigma_{\gamma(44)} \end{pmatrix}$$

This model, while very flexible, has very many parameters: I variances and $I(I-1)/2$ covariances, where I is the number of cultivars.

A further popular model for interaction is the **Finlay-Wilkinson** model, which implies regression on a latent environmental variable u_j, also known as factor, which characterized the yield potential of the j-th location. The model is

$\eta_{ij} = \mu + \alpha_i + \lambda_i u_j + \delta_{ij}$

where

λ_i = slope of *i*-th cultivar
u_j = latent variable for *j*-th location
δ_{ij} = deviation from regression of *i*-th cultivar and *j*-th location

$$u_j \sim N\left(0, \sigma_u^2\right)$$
$$\delta_{ij} \sim N\left(0, \sigma_\delta^2\right)$$

and

$$\text{var}\left(\eta_{ij}\right) = \lambda_i^2 \sigma_u^2 + \sigma_\delta^2$$
$$\text{cov}\left(\eta_{ij}, \eta_{i'j}\right) = \lambda_i \lambda_{i'} \sigma_u^2$$

The model is overparameterized, so we need to impose a restriction, e.g.:

$\sigma_u^2 = 1$.

In matrix form the model is for four cultivars:

$$\text{var}\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \\ \eta_{4j} \end{pmatrix} = \begin{pmatrix} \lambda_1^2 + \sigma_\delta^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2 + \sigma_\delta^2 & \lambda_1\lambda_2 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + \sigma_\delta^2 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2 + \sigma_\delta^2 \end{pmatrix}$$

The model allows heterogeneity both in the variances and the covariances. As the latent variable is not observable, this is truly a **factor-analytic model**, not a regression model. It is closely related to the popular regression method by Finlay and Wilkinson (1963):

(1) For each location compute the mean ($\bar{y}_{\cdot j \cdot}$) as an index for the yield potential of the *j*-th location.
(2) Yield means per cultivar and location ($\bar{y}_{ij \cdot}$) are regressed against the location mean ($\bar{y}_{\cdot j \cdot}$).

Example 20 (continued): Cultivars with good response to favorable environmental conditions show a large slope (high input cultivars), while cultivars with poor response (low input varieties) show a lower slope. An example is given in Fig. 16 for cultivars 3 and 4.

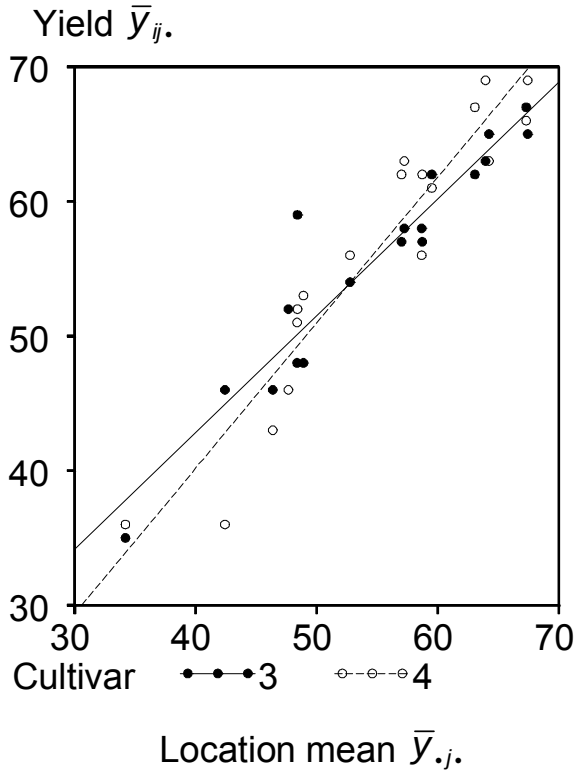


Fig. 18: Regression of cultivar-location means against location means for cultivars 3 and 4 of oats data.

Finally, consider the **Eberhart-Russell** model, which allows heterogeneity in the residual variance:

$$u_j \sim N\left(0, \sigma_u^2\right)$$
$$\delta_{ij} \sim N\left(0, \sigma_{\delta(i)}^2\right)$$

and

$$\text{var}\left(\eta_{ij}\right) = \lambda_i^2 \sigma_u^2 + \sigma_{\delta(i)}^2$$
$$\text{cov}\left(\eta_{ij}, \eta_{i'j}\right) = \lambda_i \lambda_{i'} \sigma_u^2$$

In matrix form the model is for four cultivars:

$$\text{var}\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \\ \eta_{4j} \end{pmatrix} = \begin{pmatrix} \lambda_1^2 + \sigma_{\delta(1)}^2 & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2 + \sigma_{\delta(2)}^2 & \lambda_1\lambda_2 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + \sigma_{\delta(3)}^2 & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2 + \sigma_{\delta(4)}^2 \end{pmatrix}$$

We may compare the various models by AIC and by LR tests.

Example 20 (continued): Tables 17 and 18 show fit statistics for various models as fitted to the oats data. Shukla's model (Table 19) fits best and so may be selected for final analysis.

Table 17: Fits of different models for cultivar × location interaction for oats data.

Model	AIC	$-2 \log L_R$	§ No. of variance parameters
Compound-symmetry	1307.6	1303.6	2
Skukla	1297.8	1267.8	$I + 1 = 15$
Finlay-Wilkinson	1323.8	1293.8	$I + 1 = 15$
Eberhart-Russell	1311.5	1255.5	$2I = 28$
Unstructured	no convergence$		$I(I + 1)/2 = 105$

$ Too few locations. § I = no. of cultivars.

Table 18: Comparison of nested models by LR test.

Full model	Reduced model	T	§d.f.	Critical χ^2-value $\alpha = 5\%$
Shukla	Compound symmetry	*44.2	13	22.36
Finlay-Wilkinson	Compound symmetry	10.2	13	22.36
Eberhart-Russell	Shukla	12.3	13	22.36

*significant; § difference in the number of parameters

Table 19: Variance component estimates for oats data – two-stage analysis, Shukla model.

Effect	Variance component		Estimate
Main effect location		σ^2_β	86.31
Stability variances	Cultivar 1	$\sigma^2_{\alpha\beta(1)}$	4.61
	Cultivar 2	$\sigma^2_{\alpha\beta(2)}$	14.70
	Cultivar 3	$\sigma^2_{\alpha\beta(3)}$	8.26
	Cultivar 4	$\sigma^2_{\alpha\beta(4)}$	10.51
	Cultivar 5	$\sigma^2_{\alpha\beta(5)}$	2.10
	Cultivar 6	$\sigma^2_{\alpha\beta(6)}$	0.38
	Cultivar 7	$\sigma^2_{\alpha\beta(7)}$	0.89
	Cultivar 8	$\sigma^2_{\alpha\beta(8)}$	8.26
	Cultivar 9	$\sigma^2_{\alpha\beta(9)}$	10.48
	Cultivar 10	$\sigma^2_{\alpha\beta(10)}$	3.58
	Cultivar 11	$\sigma^2_{\alpha\beta(11)}$	12.13
	Cultivar 12	$\sigma^2_{\alpha\beta(12)}$	0.97
	Cultivar 13	$\sigma^2_{\alpha\beta(13)}$	3.43
	Cultivar 14	$\sigma^2_{\alpha\beta(14)}$	5.76

There are some major differences in stability, with stability variances ranging from 0.38 (cultivar 6; relatively stable) to 14.69 (cultivar 2; relatively unstable).

Acknowledgements

I thank Andreas Büchse, Jens Möhring, Ignacio Romagosa, and Fred van Eeuwijk for critical reading of earlier versions of these lecture notes. The responsibility for any remaining errors is entirely mine.

References

Bernardo R 2002 Breeding for quantitative traits in plants. Stemma Press, Woodbury.
Bos I, Caligari P 1995 Selection methods in plant breeding. Chapman and Hall, London.
Calinski T, Kageyama S 2000 Block designs: A randomization approach. Springer, Berlin.
Calinski T, Czajka S, Kaczmarek Z, Krajewski P, Pilarczyk W 2005 Analyzing multi-environment variety trials using randomization-derived mixed models. Biometrics 61, 448-455.
Clewer AG, Scarisbrick DH 2001 Practical statistics and experimental design for plant and crop science. Wiley, New York.
Cochran WG, Cox RD 1957 Experimental designs. Wiley, New York.
Dean A, Voss D 1999 Design and analysis of experiments. Springer, New York.
Falconer DS, Mackay TFC 1996 Introduction to quantitative genetics. 4th edition. Longman, Harlow.
Gilmour AR, Cullis BR, and Verbyla AP 1997 Accounting for natural and extraneous variation in the analysis of field experiments. Journal of Agricultural, Biological and Environmental Statistics 2, 269-293.
Gilmour AR 2000 Post blocking gone too far! Recovery of information and spatial analysis in field experiments. Biometrics 56, 944-946.
Gomez A, Gomez A 1984 Statistical procedures for agricultural research. Wiley, New York.
Henderson CR 1976 A simple method for computing the inverse of a numerator relationship matrix used in predicting breeding values. Biometrics 32: 69-82.
John JA, Williams ER 1995 Cyclic and computer generated designs, Chapman and Hall, London.
Kempton, RA, Fox PN (ed) 1997 Statistical methods for plant variety evaluation. Chapman and Hall, London.
Kenward MG, Roger JH 1997 Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53, 983-997.
McCulloch CE, Searle SR 2001 Generalized, linear, and mixed models. Wiley, New York.
Mead R, Curnow RN, Hasted, AM 1993 Statistical methods in agriculture and experimental biology. Second edition. Chapman & Hall, London.
Melchinger AE, Piepho HP, Utz HF, Muminovic J, Wegenast T, Torjek O, Altmann T. Kusterer B 2007 Genetic basis of heterosis for growth-related traits in Arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. Genetics 177, 1827-1837.
Möhring, J., Melchinger, A.E., Piepho, H.P. (2011): REML-based diallel analysis. *Crop Science* **51**, 470-478.
Mrode RA 1998 Linear models for the prediction of animal breeding values. CAB International, Wallingford.
Panter DM, Allen FL 1995 Using best linear unbiased predictions to enhance breeding for yield in soybean. 2. Selection of superior crosses from a limited number of yield trials. Crop Science 35: 405-410.
Patterson HD, Hunter EA 1983 The efficiency of incomplete block designs in National List and Recommended List cereal variety trials. Journal of Agricultural Science, Cambridge 101:

427-433.
Petersen RG 1994 Agricultural field experiments. Marcel Dekker, New York.
Piepho HP 1997 Analysis of a randomized complete block design with unequal subclass numbers. Agronomy Journal 89, 718-723.
Piepho HP, Büchse A, Emrich K 2003 A hitchhiker's guide to the mixed model analysis of randomized experiments. Journal of Agronomy and Crop Science 189, 310-322.
Piepho HP, Büchse A, Richter C 2004 A mixed modelling approach to randomized experiments with repeated measures. Journal of Agronomy and Crop Science 190, 230-247.
Piepho HP, Möhring J 2006 Selection in cultivar trials – is it ignorable? Crop Science 146, 193-202.
Piepho, H.P., Möhring, J. (2011): On estimation of genotypic correlations and their standard errors by multivariate REML using the MIXED procedure of the SAS System. *Crop Science*
Piepho HP, Möhring J, Melchinger AE, Büchse A 2008a BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161, 209-228.
Piepho HP, Richter C, Williams E 2008b Nearest neighbour adjustment and linear variance models in plant breeding trials. Biometrical Journal 50, 164-189.
Piepho HP, Williams ER, Fleck M 2006 A note on the analysis of designed experiments with complex treatment structure. HortScience 41, 446-452.
SAS Institute, Inc. 1999 SAS/STAT User's Guide. SAS Institute, Cary, NC.
Schabenberger O, Pierce FJ 2002 Contemporary statistical models. CRC Press, Boca Raton.
Searle SR 1971 Linear models. Wiley, New York.
Searle SR, Casella G, McCulloch CE 1992 Variance components. Wiley, New York.
Thöni H 1970 Die Schätzung von Wachstumskurven auf Grund wiederholter Messungen am gleichen Individuum. 1. Gemischtes Modell für lineare Regression. Schweizerische landwirtschaftliche Forschung 9, 54-67.
Williams ER 1986 A neighbour model for field experiments. Biometrika 73, 279-287.
Williams ER, John JA 2003 A note on the design of unreplicated trials. Biometrical Journal 45, 751-757.
Williams ER, John JA, Whitaker D 2006 Construction of resolvable row-column designs. Biometrics 62, 103-108.
Williams ER, Fu YB 1999 Comment - enhanced heritabilities and best linear unbiased predictors through appropriate blocking of progeny trials. Can J For Res 29, 1633-1634.

A. Exercises for practicals

Some of the exercises are marked as "difficult". You may skip these at first pass through these exercises and focus on the other exercises. No solutions are provided to these exercises. For exercises labelled with a "C", solutions are provided in Section C.

Exercise 1: Reproduce all examples in the lecture notes.

Exercise 2 (Chapter 2): A trial with three treatments was laid out as an RCBD. The treatments were

O = control (natural daylight only)
E = Extended day (total day length 14 hours)
F = flash lighting (natural day + 2 x 20 second flashes per night)

The number of eggs laid by a pen of 6 pullets (young hen, less than 1 yr old) was recorded for each unit during a period of three months (Mead, Curnow and Hasted: Statistical methods in agriculture and experimental biology, Example 5.1):

	Blocks			
Treatments	1	2	3	4
O	330	288	295	313
E	372	340	343	341
F	359	337	373	302

Perform an analysis of variance, followed by mean comparisons (LSD) using PROC GLM and PROC MIXED. Compare the results.

Hints: Code the block and treatment factors as **`block`** and **`trt`**, respectively, and the response as **`number_of_eggs`**. Then use the following codes in GLM and MIXED:

```
proc glm;
class block trt;
model number_of_eggs=block trt;
means trt/lsd;
run;

proc mixed;
class block trt;
model number_of_eggs=block trt;
lsmeans trt/pdiff;
run;
```

These codes fit fixed effects for both blocks and treatments. For comparison, fit blocks as a random effect as follows:

```
proc glm;
class block trt;
model number_of_eggs=block trt;
random block;
means trt/lsd;
run;
```

```
proc mixed;
class block trt;
model number_of_eggs= trt;
random block;
lsmeans trt/pdiff;
run;
```

Notice an important difference in the use of GLM and MIXED as regards random effects: In GLM, a random effect is listed twice, i.e. in the model statement and in the random statement. By contrast, the random effect is listed only in the RANDOM statement in MIXED.

Delete the last observation from your dataset and rerun the analyses. What differences do you observe between the analyses? In particular, study the standard errors of a difference obtained from the MIXED procedure with fixed and with random block effects.

Exercise 3 (**cochran&cox.dat**) (Chapter 3): In this example, from Cochran and Cox (1957, p. 406), the data are yields (**`Yield`**) in bushels per acre of 25 varieties (**`Treatmnt`**) of soybeans. The data are collected in two replicates (**`Group`**) of 25 varieties in five blocks (**`Block`**) containing five varieties each. This is an example of a partially balanced square lattice design. Perform a mixed model analysis of this experiment.

Exercise 4 (**pbib.dat**) (Chapter 3): A partially balanced incomplete block design. Analyse this by an appropriate mixed model.

Exercise 5 (**presterl.dat**) (Chapter 3): This is data from an experiment with maize laid out as an augmented design with incomplete blocks and lots of entries. Each block is rectangular and has plots laid out in several rows and columns. The data file also contains codings for rows and columns of the field layout. Consider using these for post blocking and compare this to an analysis that considers effects of blocks, but not of rows and columns. How does the s.e.d. change when random effects are added for rows and columns. How about fixed effects for rows and columns? Would you take blocks as fixed or random here? Beware: Computing time may be long! Use the **`lognote`** option to monitor convergence progress.

Exercise 6 (Cochran & Cox, 1957, p.448) (Chapter 3): A balanced incomplete block design with 13 blocks of size four served to evaluate the performance of 13 hybrids of corn (**corn.dat**). Analyse this experiment by a suitable mixed model. Inspect the standard errors of a difference (s.e.d.). How many different values do you find? Discuss the meaning of the term "balanced" in light of your finding: In what sense is this design "balanced"?

Exercise 7 (Chapter 3) (from Dean and Voss, Design and analysis of experiments, p.370): An experiment run in 1993 investigated the effects on heart rate due to the use of a step machine. The experimenters were interested in checking the theoretical model that says that heart rate should be a function of body mass, step height, and step frequency. The experiment involved the two factors "step height" (factor C) and "step frequency" (factor D). Levels of "step height" were 5.75 and 11.5 inches, coded as 1 and 2. "Step frequency" had three equally spaced levels, 14, 21 and 28 steps per minute, coded 1, 2, 3. The response variable was pulse rate in beats per minute.

Table C2: Design and data for the step experiment

	Treatment combination					
Block	11	12	13	21	22	23
1		75	87	84	93	99
2	93	84	96	90	108	
3	99	93	96		123	129
4	99	108	99	99		120
5	99		111	90	129	141
6	129	135		120	147	153

As the table shows, this experiment was laid out in incomplete blocks. In fact, the design was a balanced incomplete block design.

Perform an analysis of this experiment both with and without recovery of inter-block information. Pay particular attention to the possible presence of interaction between the two treatment factors C and D. Determine the standard errors of treatment comparisons of interest. Can you see in the results, why the design is called “balanced”? Can you explain from the design manifest in Table C2 in what sense the design is balanced?

Exercise 8 (Chapter 3) (from Dean and Voss, p.380): Cochran and Cox (1957) describe an experiment that was run to compare the effects of cold storage on the tenderness of beef roasts. Six periods of storage (0, 1, 2, 4, 9, and 18 days) were tested and coded 1-6. It was believed that roasts from similar positions on the two sides of the animal would be similar, and therefore the experiment was run in b=15 blocks of size k=2. The response y_{ih} from treatment i in block h is the tenderness score. The maximum score is 40, indicating very tender beef. The design and responses are shown in Table C3.

Table C3: Design and data for the beef experiment

	Treatment (TIME)					
Block	1	2	3	4	5	6
I	7	17	.	.	.	.
II	.	.	26	25	.	.
II	.	.	.	.	33	29
IV	17	.	27	.	.	.
V	.	23	.	.	27	.
VI	.	.	.	29	.	30
VII	10	.	.	25	.	.
VIII	.	26	.	.	.	37
IX	.	.	24	.	26	.
X	25	.	.	.	40	.
XI	.	25	.	34	.	.
XII	.	.	34	.	.	32
XIII	11	.	.	.	.	27
XIV	.	24	21	.	.	.
XV	.	.	.	26	32	.

(1) What benefit do you think the experimenters expected to gain by using a block design instead of a completely randomized design?
(2) Carefully inspect the design and check how often each pair of treatments is in the same block. Do you, or do you not, expect this design to be variance-balanced, meaning that all pairwise contrasts are estimated with equal precision?
(3) Calculate treatment differences and their standard errors and confidence intervals
(4) Calculate the confidence interval for the difference of averages contrast

$$\frac{1}{3}(\tau_4+\tau_5+\tau_6)-\frac{1}{3}(\tau_1+\tau_2+\tau_3),$$

where τ_i is the effect of the *i*-th treatment. What does your interval tell you about storage time and tenderness of beef? You can use an estimate or a contrast statement. Consult the online help to find out how these statements are used.
(5) Why is it incorrect to estimate the difference in the effects of treatments *i* and *p* as $\bar{y}_{i\bullet}-\bar{y}_{p\bullet}$?
(6) Perform an analysis for the data in Table C3. Compare analysis with and without recovery of inter-block information. The analysis with fixed blocks yields:

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	1571.633333	82.717544	10.70	0.0003
Error	10	77.333333	7.733333		
Corrected Total	29	1648.966667			

R-Square	Coeff Var	Root MSE	y Mean
0.953102	10.84871	2.780887	25.63333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
block	14	1051.466667	75.104762	9.71	0.0005
time	5	520.166667	104.033333	13.45	0.0004

Source	DF	Type III SS	Mean Square	F Value	Pr > F
block	14	511.8666667	36.5619048	4.73	0.0090
time	5	520.1666667	104.0333333	13.45	0.0004

Parameter	Estimate	Standard Error	t Value	Pr > |t|
linear TIME	95.6666667	13.4329611	7.12	<.0001

```
The GLM Procedure
Least Squares Means

time        y LSMEAN

1         14.6333333
2         23.8000000
3         26.9666667
4         28.3000000
5         30.8000000
6         29.3000000
```

What does the output tell you? Interpret!

(7) The linear contrast is an estimate of

$$L = -5\tau_1 - 3\tau_3 - \tau_3 + \tau_4 + 3\tau_5 + 5\tau_6,$$

which is sensitive to linear trend in storage time. This contrast can be tested by an estimate statement or a contrast statement. Alternatively, one may just fit a linear regression on the treatment variable. Compare both approaches. Consult the online help on how to use the estimate and the contrast statements.

(8) How would you test for lack-of-linearity in the response (compare section 6.6 of lecture "Biometrie")?

Exercise 9 (Chapter 3)**:** Suppose you want to test six treatments in ten blocks of three units. Develop a suitable incomplete block design for this purpose. The following questions may help you to find a design that balances the number of direct comparisons for pairs of treatments.

(i) How many replications are needed for each treatment, if each is equally important?
(ii) In how many blocks will each treatment occur?
(iii) How many direct comparisons can be made of a treatment with the other treatments in the experiment?
(iv) How many direct comparisons will there be for each pair of treatments, provided the number of direct comparisons can be made to be constant for all pairs?

It may be best to try and develop the design in a text editor, using symbols A-F for the treatments in this template (first block filled for illustration):

	Block									
	I	II	III	IV	V	VI	VII	VIII	IX	X
A	.	.	.	.	.	.	.	.	.	.
B	.	.	.	.	.	.	.	.	.	.
C	.	.	.	.	.	.	.	.	.	.
D	.	.	.	.	.	.	.	.	.	.
E	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	.	.	.	.	.

[R. Mead: The design of experiments. Cambridge University Press, Cambridge, p.137 (Example 7.4)]

Exercise 10 (Chapter 3): You are required to design an experiment to compare four treatments O (control), A, B, C. The experimental material available consists of 20 units in two blocks of four units, two blocks of three units and three blocks of two units. (R. Mead: The design of experiments, p.173)

(1) Assume that block effects are fixed. Find appropriate designs to satisfy the following criteria:
(i) Variances of treatment differences between A, B and C to be as nearly equal as possible;
(ii) Variances of comparisons of O with other treatments to be approximately two-thirds those of comparisons between other treatments.

Use PROC MIXED to obtain the standard errors of all differences, assuming the error variance equals 1. Fix the residual error using the following statement:

```
parms (1)/hold=1;
```

Obtain all pairwise comparisons using this statement (assuming treatments are coded "trt"):

```
lsmeans trt/pdiff;
```

(2) Evaluate your preferred design from (1), now assuming that blocks are random and that the block variance is half that of error. Fix the block and error variances in MIXED as follows:

```
parms (0.5)(1)/hold=1,2;
```

How do the pairwise standard errors (and variances) change and why?

Exercise 11 (Chapter 3): A triple 4 x 4 lattice was used to test 16 genotypes of spring wheat (Schuster & Lochow, p.151).

Block	Replicate I				genotype number
(4)	13	14	15	16	
	4.68	4.86	4.76	4.73	← yield
(3)	9	10	11	12	
	4.89	5.25	4.46	5.13	
(2)	5	6	7	8	
	4.72	4.15	5.80	4.78	
(1)	1	2	3	4	
	5.05	5.87	4.70	5.19	

Block	Replicate II			
(8)	4	8	12	16
	5.16	4.97	4.57	4.67
(7)	3	7	11	15
	4.74	5.73	4.24	4.50
(6)	2	6	10	14
	4.96	3.29	4.00	4.32
(5)	1	5	9	13
	4.78	4.53	4.49	5.12

Block	Replicate III			
(12)	4	5	10	15
	4.78	4.61	4.58	3.76
(11)	3	8	9	14
	4.17	4.55	4.66	4.34
(10)	2	7	12	13
	4.17	4.55	5.13	4.94
(9)	1	6	11	16
	4.68	4.31	4.61	4.37

(1) Perform analysis with and without recovery of information. The textbook of Schuster and Lochow (p.156) reports this ANOVA Table for intra-block analysis:

Source	SS	d.f.	
Total	10.5853	47	
Replicates	1.5487	2	
Blocks	1.4411	9	
Treatments	5.0229	15	
Error	2.5726	21	$s^2 = 0.1225$

How does this compare to your results?

(2) Compute the efficiency relative to a randomized complete block design

(3) It appears from the layout that the design was not properly randomized. Which violations of randomization rules for lattice designs can you identify?

(4) Do you find a treatment pair with more than one concurrence (a pair that appears in the same block more than once)? Comment on your finding.

Exercise 12 (Chapter 3) (Steel and Torrie, 1980, p.164): One method of sampling fish in a lake is to kill them all by the use of rotenone, collect them in buckets, and then take a random sample of buckets. In one such experiment, a random sample of 2 buckets out of 20 was taken and all fish in each bucket measured for length in inches.

The sampling was done at three collection times in the afternoon, and one objective of the analysis was to determine, if there are significant differences between the collection times. The data are given here in tabular form and can also be found in the datafile **C7.dat**.

Collection Time		Length category (inches)							
	Sample	3	4	5	6	7	8	9	*n*
1 : 50	A	.	.	5	19	19	8	3	54
	B	.	.	10	27	15	6	3	61
3 : 20	A	.	4	11	26	10	11	3	65
	B	.	3	11	29	13	8	1	65
4 : 40	A	2	8	16	44	15	8	.	93
	B	1	6	15	35	12	7	2	78

Define a suitable mixed model for analysis. Test the null hypothesis that collection times do not differ in the size of fish caught. Perform pairwise comparisons among collection times using a t-test.

Exercise 13 (Chapter 3) (Schuster & Lochow, 1979, p.171):

A two-factorial experiment was performed to study the effect of three sowing times with three different cruciferous plant species. Due to technical reasons a split-plot design was used, where sowing times were randomized on main plots and species on sub-plots. The main plots were laid out in complete blocks, while sub-plots were completely randomized within main-plots. Plot yields (kg fresh weight per plot) were as follows.

d	1 31,2	I 3 10,5	2 15,6	3 13,8	III 2 9,8	1 10,5	2 24,2	II 1 26,8	3 14,8
c	2 8,5	III 3 15,3	1 9,8	3 8,5	I 1 28,9	2 13,0	1 24,3	II 2 25,5	3 14,2
b	3 18,2	II 1 29,2	2 28,1	1 12,0	III 2 13,0	3 18,5	2 17,3	I 3 12,3	1 35,2
a	1 30,5	I 2 15,1	3 10,0	2 27,1	II 3 26,9	1 15,5	3 10,3	III 1 12,5	2 17,6

Blocks are labels a-d, sowing times I-III, and species 1-3. Perform a mixed model analysis of this experiment, including suitable mean comparisons following the analysis of variance (F-tests). Use the Kenward-Roger methods to determine the denominator degrees of freedom.

The textbook (Schuster & Lochow, 1979, p.171) gives these F-statistics:

Source	d.f.	F-statistic
Sowing times	2	16.7731
Species	2	11.2145
Sowing times × species	4	11.6689

Can you verify these results? (textbooks are not always right!)

Exercise 14 (Steel and Torrie, 1980:154) (Chapter 4): A 3 x 2 factorial experiment was performed with mint plants as a completely randomized design with three replicates and four sub-samples (plants) per experimental unit (pot). From a large group of plants four were randomly assigned to each of 18 pots. Six treatments were then randomly assigned to pots. The treatments consisted of all combinations of three hours of daylight (8, 12 and 16 hrs) and two levels of night temperatures (high, low). Stem growth was recorded for the mint plants (**mint.dat**). Perform an analysis of variance accounting for the subsampling structure.

Low night temperatures

	8 hrs			12 hrs			16 hrs		
Plant	Pot 1	Pot 2	Pot 3	Pot 1	Pot 2	Pot 3	Pot 1	Pot 2	Pot 3
1	3.5	2.5	3.0	5.0	3.5	4.5	5.0	5.5	5.5
2	4.0	4.5	3.0	5.5	3.5	4.0	4.5	6.0	4.5
3	3.0	5.5	2.5	4.0	3.0	4.0	5.0	5.0	6.5
4	4.5	5.0	3.0	3.5	4.0	5.0	4.5	5.0	5.5

High night temperatures

	8 hrs			12 hrs			16 hrs		
Plant	Pot 1	Pot 2	Pot 3	Pot 1	Pot 2	Pot 3	Pot 1	Pot 2	Pot 3
1	8.5	6.5	7.0	6.0	6.0	6.5	7.0	6.0	11.0
2	6.0	7.0	7.0	5.5	8.5	6.5	9.0	7.0	7.0
3	9.0	8.0	7.0	3.5	4.5	8.5	8.5	7.0	9.0
4	8.5	6.5	7.0	7.0	7.5	7.5	8.5	7.0	8.0

Exercise 15 (Peterson, 1994, p. 136) (Chapter 5): "A breeder wanted to determine the effect of planting date on the yield of four varieties of winter wheat. She chose two factors as follows:

Planting date: D1 = Oct. 15, D2 = Nov. 1; D3 = Nov. 15
Variety: V1; V2; V3; V4

To facilitate the operations of planting and harvesting she decided to use a split-plot design with planting dates assigned at random to strips of plots within each of three blocks. Varieties were then assigned at random to plots within planting dates. "

Analyse the data in **split plot date.dat**. Determine if there is a linear time trend with respect to planting date.

Exercise 16 (Chapter 5; taken from Clewer and Scarisbrick 2002, p.242): An experiment was carried out to find the effect of spraying an insecticide to control aphids on the yield of three varieties of field bean. Factor A was spraying at two levels: A_1 = no insecticide (plots sprayed with water), A_2 = + insecticide in the same volume of water. Factor B was variety with B_1 = variety 1, B_2 = variety 2 and B_3 = variety 3. The field layout and yields are shown below.

a. What kind of a design is this?
b. Perform an analysis of variance based on a suitable linear model. The data are found in **spray.dat**. Briefly summarize the results.
c. Which kind of mean comparisons is suitable? Justify your choice based on the outcome of the ANOVA.

Block I

25.5	24.9	25.8	26.1	18.0	21.7
A_2B_2	A_2B_1	A_2B_3	A_1B_3	A_1B_2	A_1B_1

Block II

21.1	17.9	28.9	27.6	29.4	29.3
A_1B_1	A_1B_2	A_1B_3	B_1 A_2	B_3 A_2	B_2 A_2

Block III

28.6	19.5	23.2	29.7	30.3	29.5
A_1B_3	A_1B_2	A_1B_1	B_2 A_2	B_3 A_2	B_1 A_2

Block IV

29.3	29.2	29.8	21.3	31.0	25.8
B_3 A_2	B_2 A_2	B_1 A_2	A_1B_2	A_1B_3	A_1B_1

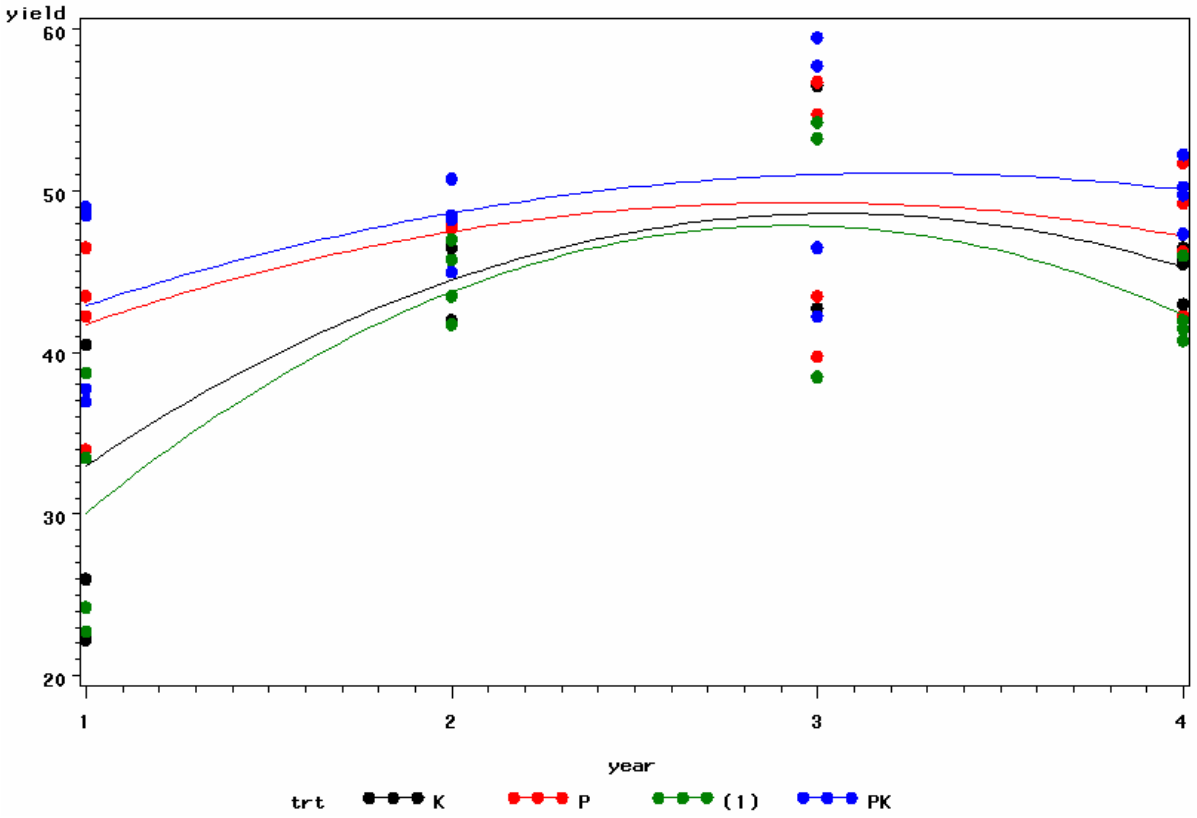
Exercise 17 (Peterson, 1994, p.277) (Chapter 6): "An agronomist was interested in examining the response of alfalfa to the application of phosporus and potassium fertilizers. She was also interested in observing the nature of response to fertilizer as a function of time after application. She established an experimental planting of alfalfa using a randomized complete block design with four blocks of four plots each. The fertilizer treatments consisted of the following 2 x 2 factorial combinations of phosphorus and potassion:

(1) = no fertilizer
P = 20 kg/ha phosphorus

K = 20 kg/ha potassion
PK = 20 kg/ha P and 20 kg/ha K

Ferlilizer was applied in the spring of the year following the establishment of the trial. The plots were clipped and weighed three times each year for the duration of the trial. Plot yields were then expressed as total for the three cuttings per year. The trial was conducted for four years. After which it was abandoned."

The data is stored in **PK.dat**. Analyse this trial by a suitable mixed model. Do time profiles differ among treatments? Is there a linear or quadratic trend? Test the lack of fit four your time trend model.



Exercise 18 (Schabenberger & Pierce, 2002, p.466) (Chapter 6): "At the Winchester Agricultural Experiment Station of Virginia Polytech Institute and State University ten apple trees were randomly selected and twenty-five apples were randomly chosen on each tree. We concentrate the analysis on the apples in the largest size class, those whose diameter exceeded 2.75 inches. In total there were 80 apples in the largest size class. Diameters of the apples were recorded in two-week intervals over a twelve-week period" (**AppleData.xls**). Fit a mixed model to assess the overall time trends. The model should reflect the two-stage sampling scheme (trees at first stage, apples within trees at second stage).

(i) Find a suitable variance-covariance model for modelling the autocorrelation of diameter measurements taken on the same apples. Specifically, consider these models: CS, UN, AR(1), ARH(1), CSH, independent with heterogeneous variance (UN(1)) and independent.
(ii) Using the selected variance-covariance model, find a suitable polynomial regression model for the diameter growth over time.

(iii) Once this has been selected, test for heterogeneity among regression curves of the different trees.
(iv) Report the final model.

(In case a regression model does not converge, try rescaling the time regressor variable. For example, use t = time – 5 as a regressor variable instead of time).

Exercise 19 (Chapter 6): The data of Example 12 (Lemna data; section 4.6; **lemna.dat**) were log-transformed counts. The transformation was used because it stabilized the variance across the three time points. The original counts are stored in the variable **`count`**. Repeat the analysis of this repeated measures data with the response variable count. Which covariance structure do you select? Does the result for the time trend regression change?

Exercise 20 (Chapter 6): Explore the effect of a log transformation of rosette diameter for the Arabidopsis data in Section 4.7.3 (**arabidopsis.xls**). Which variance-covariance structure fits well for the transformed data?

SAS-Hint: The rosette diameters for the three time points are stored as three variables (RD15m RD22, RD29) in a single record per rat. For analysis by MIXED, this needs to be re-arranged as three records per rat, one for each time point. A single variable y needs to be generated, plot a time variable. After importing the Excel file using the import facility, you can use this code to generate a new dataset e7.

```
/*use import facility to read arabidopsis.xls into a*/

data e7;
set a;
y=RD15; time=15; output;
y=RD22; time=22; output;
y=RD29; time=29; output;
run;

proc print data=e7; run;
```

Exercise 21 (Chapter 6): Write down the covariance structure for the ARH(1) model (a repeated measures covariance model) in the form $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}$ for three time points and a single subject. Thus, $\boldsymbol{D}$ and $\boldsymbol{C}$ are 3 x 3 matrices. $\boldsymbol{C}$ is the correlation matrix and $\boldsymbol{D}$ a diagonal matrix of time-point-specific variances. Write down $\boldsymbol{D}$ and $\boldsymbol{C}$ and then detrive $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}$. Verify that

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho^2\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 \end{pmatrix}$$

Exercise 22 (Chapter 6): Two different feed additives (Thiouracil, Tyroxin) and a control were tested on 10, 7 and 10 rats, respectively (**rats.dat**). The layout of the design was completely randomized, i.e., treatments were randomly allocated to rats. The weight of the animals was recorded at five successive times (at beginning and in weeks 1, 2, 3 and 4 after the beginning of the feeding experiment). The objective of the experiments was to compare the growth profiles for the three treatments.

Table C9: Weight data for rats under three different treatments

	Week				
	0	1	2	3	4
Control					
Rat 1	57	86	114	139	172
Rat 2	60	93	123	146	177
Rat 3	52	77	111	144	185
Rat 4	49	67	100	129	164
Rat 5	56	81	104	121	151
Rat 6	46	70	102	131	153
Rat 7	51	71	94	110	141
Rat 8	63	91	112	130	154
Rat 9	49	67	90	112	140
Rat 10	57	82	110	139	169
Tyroxin					
Rat 1	59	85	121	156	191
Rat 2	54	71	90	110	138
Rat 3	56	75	108	151	189
Rat 4	59	85	116	148	177
Rat 5	57	72	97	120	144
Rat 6	52	73	97	116	140
Rat 7	52	70	105	138	171
Thiouracil					
Rat 1	61	86	109	120	129
Rat 2	59	80	101	111	126
Rat 3	53	79	100	106	133
Rat 4	59	88	100	111	122
Rat 5	51	75	101	123	140
Rat 6	51	75	92	100	119
Rat 7	56	78	95	103	108
Rat 8	58	69	93	114	138
Rat 9	46	61	78	90	107
Rat 10	53	72	89	104	122

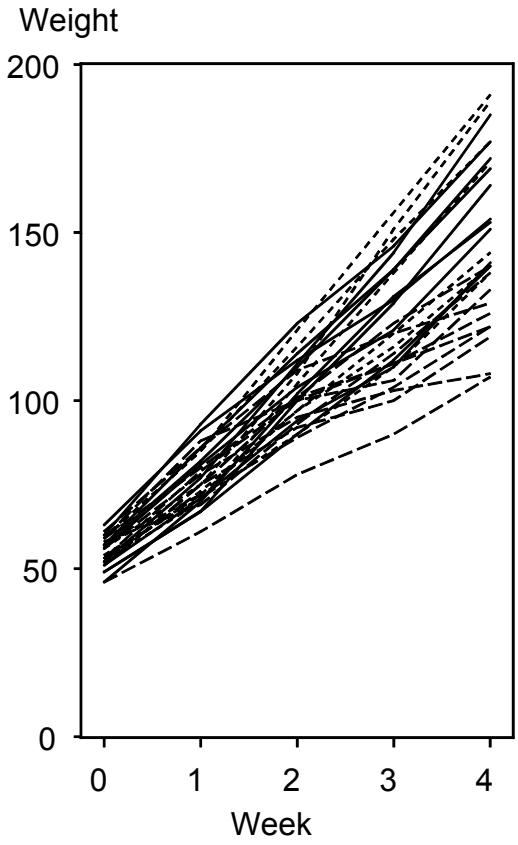


Fig. C9: Time profiles of 27 rats under three different treatments
Control ━━━ Tyroxin ▪▪▪▪▪▪ Thiouracil ▬ ▬

The profiles in Figure C9 show that there is little weight difference at the beginning, but that more pronounced differences are found at latter stages.

(1) Fit a mixed model with to study the interaction of time and treatment. Fit a suitable covariance model to account for serial correlation of observations on the same rat. In particular, compare the fits of these three models: compound symmetry, AR(1), and unstructured. Which model do you prefer?

(2) Based on the covariance model selected in (1), perform an analysis of variance, regarding time as a qualitative factor. Is there a significant interaction of treatment and time?

(3) Using the worst fitting covariance model from (1), fit a linear regression and test for lack of fit. In case there is no significant lack of fit, test whether there are differences among treatments in terms of the slopes for regression on time. Report an equation for your final regression model.

Use the best fitting covariance model from (1) and repeat the analysis. What differences do you observe?

Exercise 23 (*difficult exercise) (Chapter 8): An association mapping study was performed to study the genetics of senescence in potatoes (**potatoes phenotypes.xls**). A score for senescence was assessed at several time points. For each genotype, a single time series is available. In addition, marker data are available (**potatoes markers.xls**). Check if any of the markers is associated with senescence. Account for the repeated-measures nature of the data (You may also use the dataset **potatoes phenotypes & markers.xls** for convenience).

Exercise 24 (*difficult exercise) (Chapter 8): The maize data of Example 17 is stored in **yield0_bivariate.dat** with two traits: yield and KTS (seed dry mass in %). The pedigree is in **pedigree0.dat**. Try a bivariate analysis using the pedigree.

Exercise 25 (**schrag.dat**) (Chapter 8): A factorial crossing experiment was conducted with two pools of maize (flint and dent). The experiment way laid out as a lattice design. Five standards were tested along with crosses. The data file also contains information on the crossing parents. In each cross, one parent is from the one pool (parent1) and one from the other (parent2). A factorial may be analysed by fitting a two-way model to the crosses:

$$\tau_{ij} = (gca)_i + (gca)_j + (sca)_{ij}$$

where $(gca)_i$ is the general combining ability (g.c.a.) of the *i*-th parent (**parent1**) from the one pool, $(gca)_j$ is the general combining ability of the *j*-th parent (**parent2**) from the other pool, and $(sca)_{ij}$ is the specific combining ability (s.c.a.) for the *ij*-th cross. Estimate variance components for these three effects. Compute BLUPs for the s.c.a. effects. Exclude the standards from the random part of the model using the variable **standard** and a suitably defined dummy variable **w** as follows:

entry_nr	standard	w
1	HELIX	0
2	SYMPHON	0
3	ATTRIBUT	0
4	TUERKIS	0
5	CLARICA	0
6-65	NA	1

The variable **yield** holds the yield per plot. **GPF** is the number of harvested plants, while **SPF** (= 52) is the number of plants originally sown per plot. On some plots **GPF** < **SPF**, indicating that some plants did not reach maturity. The plants remaining on a plot can partly make up for the gaps, but still yield on plots with missing plants will tend to be smaller than on plots with no missing plants. In order to adjust for the missing data, **yield_adj** was computed by setting the yield of the missing plants equal to 50% of the yield per plant for the observed plants and then computing the sum across all plants. Alternatively, one may analyse unadjusted yield data using **GPF** as a covariate (analysis of covariance). Adding **GPF** as a regression term to the fixed part of the model will provide effect estimates adjusted for the covariate. Add **GPF** as a covariate and compare the resulting analysis to that without the covariate.

Exercise 26 (*difficult exercise) (Chapter 8): A complete diallel of 21 Triticale genotypes was tested in five locations (**triticale.xls**). Trials were laid out as α-designs. In addition to cross performance, the *per se* performance of parents as well as of selfs was tested. *Per se* performance and that of selfs can be regarded as identical. Lines were tested in two seeding densities. Furthermore, a few check varieties were included in the trials. One objective of the study was to estimate the variance components for g.c.a. and for s.c.a. of the diallel and the BLUPs of genetic effects.

The coding of treatment factors found in the dataset is as follows:

seed_density:	1,2; codes the two seed densities
check:	1-4=standards; 101=hybrids; 102=lines
hybrid:	running number for 210 hybrids; for checks and lines **hybrid=1**
lines:	running number for lines; for hybrids and checks **lines=1** Note that lines are tested in two seed densities
parent1, parent2	Parents of hybrids (and lines=selfs); for checks **parent=1**
dummy_hybrid	1=hybrid, 0=otherwise
dummy_line	1=line, 0=otherwise

Fit a mixed model that models lines and checks as fixed effects, whily hybrids are random. Start by fitting a simple random effect for hybrids. Then try to extend the model to fit g.c.a. and s.c.a. effects. To block out selfs, *per se* tests and checks, use dummy coding as suggested in Exercise 5.

Exercise 27 (Chapter 9): A series of 8 experiments was performed with 17 fodder beet genotypes (**febeet88.dat**). Trials were laid out in complete blocks with four replicates. Analyse this dataset using a mixed model accounting for heterogeneity in error variance. Also, try fitting different models for the genotype-environment interaction component. Note that the number of genotypes is larger than that of locations, so fitting complex models will be difficult. Probably you can't go further than compound symmetry and stability variance, perhaps Finlay-Wilkinson. With SAS I am getting this result for the Finlay-Wilkinson model:

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
block	loc		7.2056
FA(1)	loc		14.5417
FA(1,1)	loc		13.8324
FA(2,1)	loc		12.9663
FA(3,1)	loc		13.7726
FA(4,1)	loc		13.8896
FA(5,1)	loc		11.2811
FA(6,1)	loc		14.0675
FA(7,1)	loc		11.4183
FA(8,1)	loc		11.4905
FA(9,1)	loc		9.9484
FA(10,1)	loc		14.9071
FA(11,1)	loc		14.5464
FA(12,1)	loc		13.7785
FA(13,1)	loc		13.3666
FA(14,1)	loc		11.1485
FA(15,1)	loc		10.3880
FA(16,1)	loc		12.6542
FA(17,1)	loc		15.2758
Residual		loc 1	32.1789
Residual		loc 2	19.6504
Residual		loc 3	30.3108
Residual		loc 4	20.8160
Residual		loc 5	64.4864
Residual		loc 6	26.9506
Residual		loc 7	13.1653
Residual		loc 8	6.2297

The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	3424.6
AIC (smaller is better)	3478.6
AICC (smaller is better)	3481.6
BIC (smaller is better)	3480.8

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
gen	16	112	15.94	<.0001

Also, try fitting models with random genotypes and fixed locations, in other words, reverse the role of genotypes and environment. Can you fit the unstructured model? It works in SAS, taking block effects fixed:

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
Var(1)	gen		167.21
Var(2)	gen		66.1500
Var(3)	gen		91.9144
Var(4)	gen		104.12
Var(5)	gen		228.57
Var(6)	gen		58.4169
Var(7)	gen		71.5295
Var(8)	gen		67.8965
Corr(2,1)	gen		0.9703
Corr(3,1)	gen		0.9579
Corr(3,2)	gen		0.9694
Corr(4,1)	gen		0.8505
Corr(4,2)	gen		0.9115
Corr(4,3)	gen		0.8858
Corr(5,1)	gen		0.8890
Corr(5,2)	gen		0.9117
Corr(5,3)	gen		1.0000
Corr(5,4)	gen		0.9148
Corr(6,1)	gen		0.8761
Corr(6,2)	gen		0.9408
Corr(6,3)	gen		0.9411
Corr(6,4)	gen		0.9834
Corr(6,5)	gen		0.9206
Corr(7,1)	gen		0.9452
Corr(7,2)	gen		0.9449
Corr(7,3)	gen		0.9805
Corr(7,4)	gen		0.9466
Corr(7,5)	gen		0.9840
Corr(7,6)	gen		0.9475
Corr(8,1)	gen		0.6885
Corr(8,2)	gen		0.7673
Corr(8,3)	gen		0.6114
Corr(8,4)	gen		0.7595
Corr(8,5)	gen		0.7551
Corr(8,6)	gen		0.6091
Corr(8,7)	gen		0.6588
Residual	gen	loc 1	31.1806
Residual	gen	loc 2	20.3076
Residual	gen	loc 3	30.2108
Residual	gen	loc 4	21.4958
Residual	gen	loc 5	58.0619
Residual	gen	loc 6	28.0746

```
Residual      gen        loc 7     13.2415
Residual      gen        loc 8      6.0588


                The Mixed Procedure

                  Fit Statistics

   -2 Res Log Likelihood          3266.1
   AIC (smaller is better)        3352.1
   AICC (smaller is better)       3360.2
   BIC (smaller is better)        3387.9
```

Exercise 28: This example concerns a line-source sprinkler-irrigation experiment, which is described in the SAS/STAT manual (SAS Institute, Inc. 1999, p. 2213): "Three cultivars (…) of winter wheat are randomly assigned to rectangular plots within each of three blocks (BLOCK). The nine plots are located side-by-side, and a line-source sprinkler is placed through the middle [of each plot]. Each plot is subdivided into twelve subplots, six to the north of the line-source, six to the south (DIR). The two plots closest to the line-source represent the maximum irrigation level (IRRIG = 6), the two next closest plots represent the next-highest level (IRRIG = 5), and so forth." Cultivars will be coded by CULTIVAR, while plots are coded by the variable ROW. The data is in **line_source_sprinkler_irrigation.dat**.

Fit a suitable mixed model that reflects the lack of randomization of irrigation levels. Is there a significant interaction between cultivar and irrigation level? What type of mean comparisons do you suggest? Is a regression analysis useful and if so, perform this.

Further details on this analysis may also be found in Piepho et al. (2004).

Exercise 29 (taken from Clewer and Scarisbrick, 2001, p.253-254): An experiment was carried out to compare the effects of three concentrations of a chemical seed dressing with a control on the yield of oats. Three varieties were used in the trial because it was suspected that the response to the seed treatment would depend on the variety used. A split plot design was laid out in five randomised blocks. Varieties were assigned at random to the main plots within each block, and the seed treatment and control were assigned at random to the subplots within each main plot. The design and yields in kg per subplot were as follows (**Clewer & Scarisbrick oats.dat**):

A_1 = Variety 1 A_2 = Variety 2 A_3 = Variety 3
B1 = Control B2 = Conc 1 B3 = Conc 2 B4 = Conc3

	Block 1		Block 2		Block 3		Block 4		Block 5
	B1 42.9		B2 67.3		B3 41.4		B2 46.3		B3 54.1
A1	B4 44.4	A3	B3 65.3	A2	B4 44.1	A1	B4 34.7	A3	B1 56.5
	B3 49.5		B1 65.6		B2 42.4		B3 39.4		B2 60.2
	B2 53.8		B4 69.4		B1 45.4		B1 30.8		B4 57.2
	B3 59.8		B2 69.6		B1 28.9		B2 58.5		B4 34.0
A2	B1 53.3	A2	B3 65.8	A1	B3 40.7	A3	B3 51.0	A1	B2 41.2
	B2 57.6		B4 57.4		B4 28.3		B4 47.4		B1 35.1
	B4 64.1		B1 69.6		B2 43.9		B1 52.7		B3 37.4
	B3 68.8		B3 53.8		B1 54.0		B3 45.4		B4 51.8
A3	B1 75.4	A1	B2 58.5	A3	B4 56.6	A2	B4 51.6	A2	B3 50.3
	B2 70.3		B1 41.6		B2 57.6		B1 35.1		B2 52.7
	B4 71.6		B4 41.8		B3 45.6		B2 51.9		B1 48.3

Analyse these data as fully as possible and state your conclusions as to the effect of the concentration of seed dressing on yield for each variety. First regard concentration of the seed dressing as a qualitative factor. Compare this to an analysis that treats concentration as a quantitative factor. Assume that the three concentrations are 50, 100 and 150 units.

Exercise 30: The layout given below is from a field trial testing five varieties of wheat (A-E). Plots are numbered 1-20 on the field map (top left corner of each plot). Yields are given to the right of the treatment labels. The layout was according to a completely randomized design.

Perform a one-way analysis of variance (ANOVA) for this trial. Use the procedures GLM, MIXED and GLIMMIX for this purpose. Compare the results. Conduct an F-test of the global null hypothesis that all varieties have equal mean. Perform all pairwise comparisons among variety means using a t-test. Compute the least significant difference (LSD).

1 B 21	2 D 34	3 D 32	4 E 24
5 C 27	6 E 23	7 A 31	8 B 23
9 A 32	10 C 29	11 E 27	12 A 37
13 D 31	14 D 27	15 A 32	16 B 25
17 B 19	18 C 34	19 E 26	20 C 34

Exercise 31: A greenhouse experiment was conducted to compare six different fertilizer treatments on the thousand kernel weight of a barley variety. There were four blocks, each

with three pots. A pot contained four plants and the mean thousand kernel weights in g per pot were assessed.

Table: Design and mean thousand kernel weight per pot

Replicate	Block	Treatments (T1 to T6) and mean thousand kernel weight (g)					
1	1	T4	61.2	T6	59.2	T3	60.2
1	2	T1	44.5	T5	49.3	T2	52.7
2	3	T5	49.5	T1	52.7	T3	48.2
2	4	T4	49.7	T6	55.5	T2	46.9

(a) Perform an intra-block analysis of this experiment. Is there a significant treatment effect? Perform a multiple comparison of means. Inspect the standard errors of a difference (SED). How many different SED do you find? Can you explain your observation? Which comparisons are significant at the 5% level?

(b) Perform an analysis with recovery of inter-block information. Is there a significant treatment effect? Perform a multiple comparison of means. Which comparisons are significant at the 5% level?

(c) Is it worthwhile in this experiment to exploit the inter-block information? How can you decide on this question?

(d) Is the design "resolvable"? Explain your answer.

Exercise 32 (Mead et al., 2012, p.155): An experiment to examine preferences of cabbage root flies for six different brassica species on which to lay their eggs involved the use of ten cages of flies with plants of three brassica species available in each cage. The number of eggs laid on the various plants were as shown in the following table.

	Species					
Cage	Cabbage	Cauliflower	Broccoli	Sprouts	Kale	Romanesco
1	452	69	83			
2	802	143		53		
3	699		32		4	
4	1207			19		32
5	958				8	8
6		328	147			53
7		314		264	223	
8		158			36	5
9			117	14	115	
10			23	16		2

Analyse this experiment by a suitable linear model and perform pairwise comparison of species means. Critically inspect the residuals for any departures from the "usual" assumptions. Consider transformations of the data such as the square root and the logarithm to improve the distributional properties of the data. Hint: Check section B.9 for instructions how to generate residuals.

Exercise 33: A field experiment with two factors was conducted to study the effect of strip irrigation and of the choice of sorghum variety on yield. The irrigation treatments were randomized in complete blocks. Each experimental unit of this design was subdivided into four subunits. There where four varieties (labelled here as A-D), which were completely randomized among subunits within a unit of the design for the irrigation treatments. The amount of irrigation water applied per plot was varied between 0 liters and 1600 liters. For all irrigation treatments, the frequency of irrigation across the vegetation period was identical. Only the total amount of irrigation water varied. The data is stored in **irrigation.dat**.

(a) What is the common name for this randomization layout?
(b) How many irrigation treatments were tested in this experiment?
(c) Perform an analysis to assess the effect of irrigation and variety on yield. Select an appropriate model by suitable tests and justify your model choice.
(d) Based on the selected model, is there a significant effect of irrigation and of variety?
(e) Report the final result in a suitable form.

Exercise 34: Repeat Exercise 33, but use the data in **irrigation2.dat**.

Exercise 35: This example concerns a trial with three soil preparation methods (SOIL) and four soil depths (DEPTH). The trial was laid out in randomized complete blocks. The amount of mineral nitrogen (Nmin) was measured at four soil depths (DEPTH). The data were kindly provided by Fabian Wald (Institut für Pflanzenbau und Grünland, Universität Hohenheim, Germany). The dataset is stored in **wald.dat**. Fit a model to study the effect of depth and soil preparation method on nitrogen level. Carefully check the residuals for any departure from the usual assumptions. If a departure is detected, consider a suitable data transformation.

Exercise 36: (a) Consider the polynomial regression for the split-plot example in Chapter 5. We fitted a quadratic regression. If you inspect the plots and the estimated regression equation, you'll find that the quadratic term for variaty V1 is relatively small. Verify that this quadratic term is not significant. Try and fit a model with a quadratic response for the varieties V2-V4 and a linear response for V1. Test the null hypothesis that the linear term is not significant. Hint: Define a covariate SWTCH that can be used to switch off the quadratic term for variety V1. This variable takes the value SWITCH=0 for V1 and SWITCH=1 otherwise. To use this switch variable, simply cross it with the quadratic term in the model statement:

```
v*n_amount*n_amount*switch
```

Try this trick and inspect the solution. For more details on this trick see

Piepho, H.P., Williams, E.R., Fleck, M. (2006): A note on the analysis of designed experiments with complex treatment structure. *HortScience* **41**, 446-452.

(b) Consider the quadratic model again. Test the null hypothesis that the curves coincide for V2 and V3. You can use a CONTRAST statement for this purpose in which you simultaneously specify the equality of intercept, linear and quadratic term for these two varieties. The three equality hypotheses need to be specified in a single CONTRAST statement and separated by commas. For the underlying theory of constructing hypotheses in linear models, consult Secion 4.2 and in particular the specification of the hypothesis matrix ***K***. For exact usage of the CONTRAST statement, consult the online documentation for the

MIXED procedure. This analysis should lead to an F-statistic with three numerator degrees of freedom. This test is known as “test of coincidence”. For details see the book chapter

Richter, C., Piepho, H.P. (2015): Linear regression techniques: In: Book edited by Barry Glaz et al. American Society of Agronomy (forthcoming)

(c) Perform a test of conincidence for all four varieties simultaneously based on the quadratic model. Convince yourself, that the test has nine numerator degrees of freedom and that there are several ways to specify the null hypothesis in a single CONTRAST statement. Again see Section 4.2 for some relevant background.

B. A short MIXED manual

All documentation for the MIXED procedure can be found online using the help menu of SAS. For this reason, a separate manual is not really needed. However, for ease of first usage, the most important features needed for this lecture are described here for the different statements.

1. PROC MIXED call

PROC MIXED <options>;

There are many options which can be used with the call of mixed.

DATA=<mydata> This statement uses the file <mydata> for analysis.

LOGNOTE This option gives you a report of the iteration progress in the log-window. Use of this option is particularly useful if you have a big model to fit and expect long computing time

METHOD=<method> This option allows you to change the estimation method for variance components. The default method is METHOD=REML, so you do not need to specify this. The ML method is specified by METHOD=ML.

2. CLASS statement

Here you list all factors to be used in the mixed model that are qualitative factors or classification factors. If you use such a factor to specify a model term, then a separate effect is fitted for each level of the factor.

Quantitative factors, on which a regression is to be performed, should NOT be listed in CLASS statement!

3. MODEL statement

MODEL <depvar>=<effects>/<options>;

<depvar> is your dependent variable.
<effects> are all fixed effects in the model.

The following options are important:

DDFM=Satterth Specifies the Satterthwaite method for approximating the denominator degrees of freedom.

DDFM=KR Specifies the Kenward-Roger method for approximating the denominator degrees of freedom. This differs from the Satterthwaite method in that a variance inflation due to estimation of variance

components is taken into account.

SOLUTION Prints solution of fixed effects

4. RANDOM statement

RANDOM <effects>/<options>;

Here you list the random effects. An important option is as follows:

SOLUTION Prints solution of random effects (BLUPs)

5. REPEATED statement

This statement can be used to model the residual error term. It is particularly useful, when there are serially correlated measurements taken on the same unit (subject). For example, to specify an AR(1) model for measurements at different times (TIME) taken on the same animal (ANIMAL) the statement is

REPEATED time/SUBJECT=animal TYPE=AR(1);

6. LSMEANS statement

LSMEANS <effect>/<options>;

This statement allows you to compute adjusted means. You need to specify the effect for which you want to compute means. All terms appearing in an effect used with this statement must have appeared previously in the CLASS statement, and the same effect must also appear in the MODEL statement.

The following options are frequently needed:

DIFF Computes all pairwise differences with standard errors

PDIFF Computes p-values for pairwise t-tests

ADJUST=<method> This allows you to adjust for multiple testing in order to control the family-wise Type I error rate. For example, ADJUST=TUKEY selects the Tukey test.

7. ODS tables

You can route various components of the output produced by the MIXED procedure to a SAS dataset via the Output Delivery System (ODS). The manual tells you which datasets can be generated. Two important ones are:

DIFFS Contains all pairwise differences with tests. Requires that the option DIFF or PDIFF be used with the LSMEANS statement.

LSMEANS Contains the adjusted means. Requires that the LSMEANS statement is used.

The following ODS statement generates datasets LSMEANS and DIFFS:

ODS OUTPUT DIFFS=DIFFS LSMEANS=LSMEANS;

This statement must be placed directly before the call of PROC MIXED, or it can be integrated as an extra line into the MIXED statements.

8. ODS graphics

You can generate a number of residual plots for mixed model analysis by using this statement before a call of PROC MIXED:

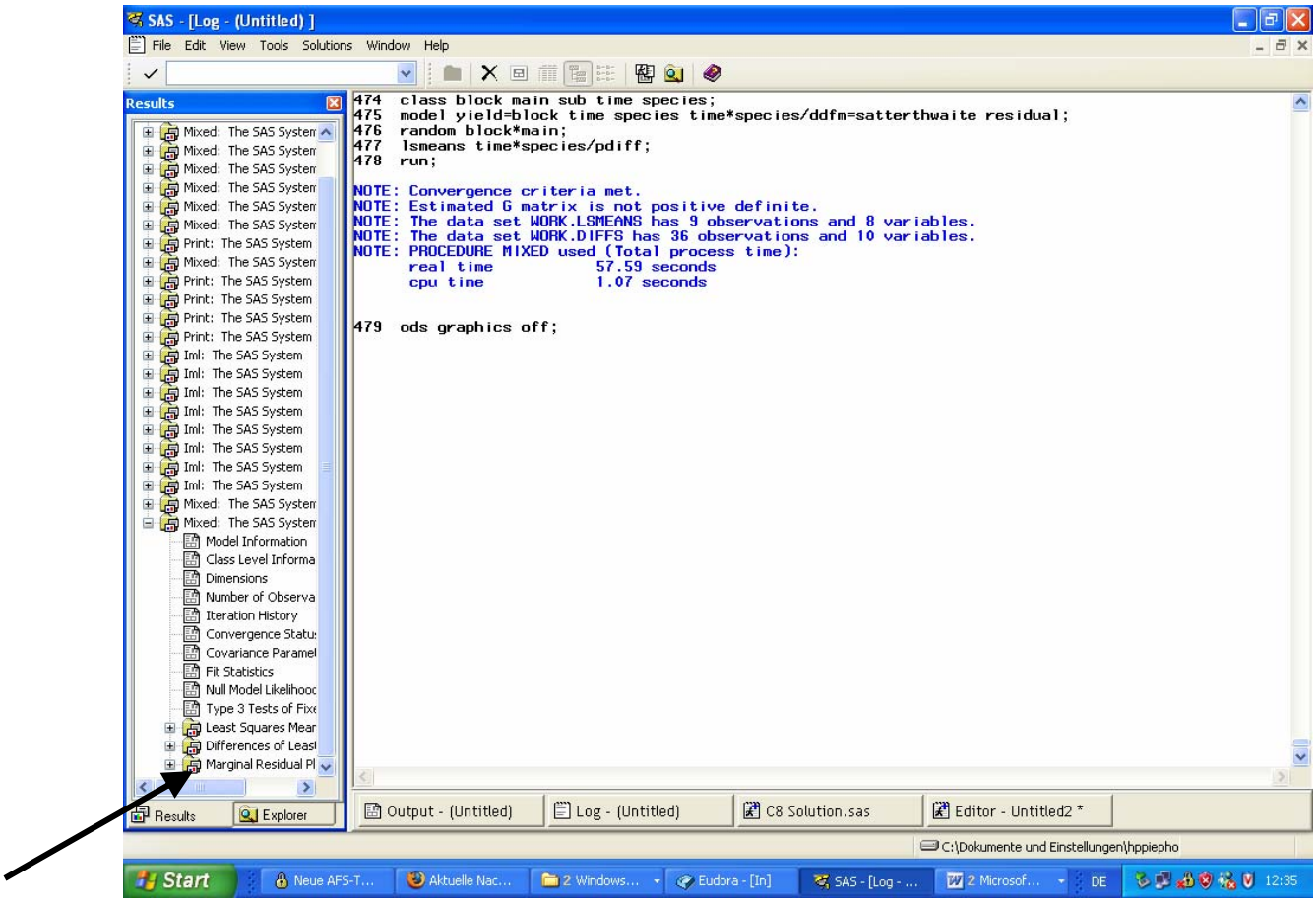
ODS GRAPHICS ON;

After the procedure, use this statement:

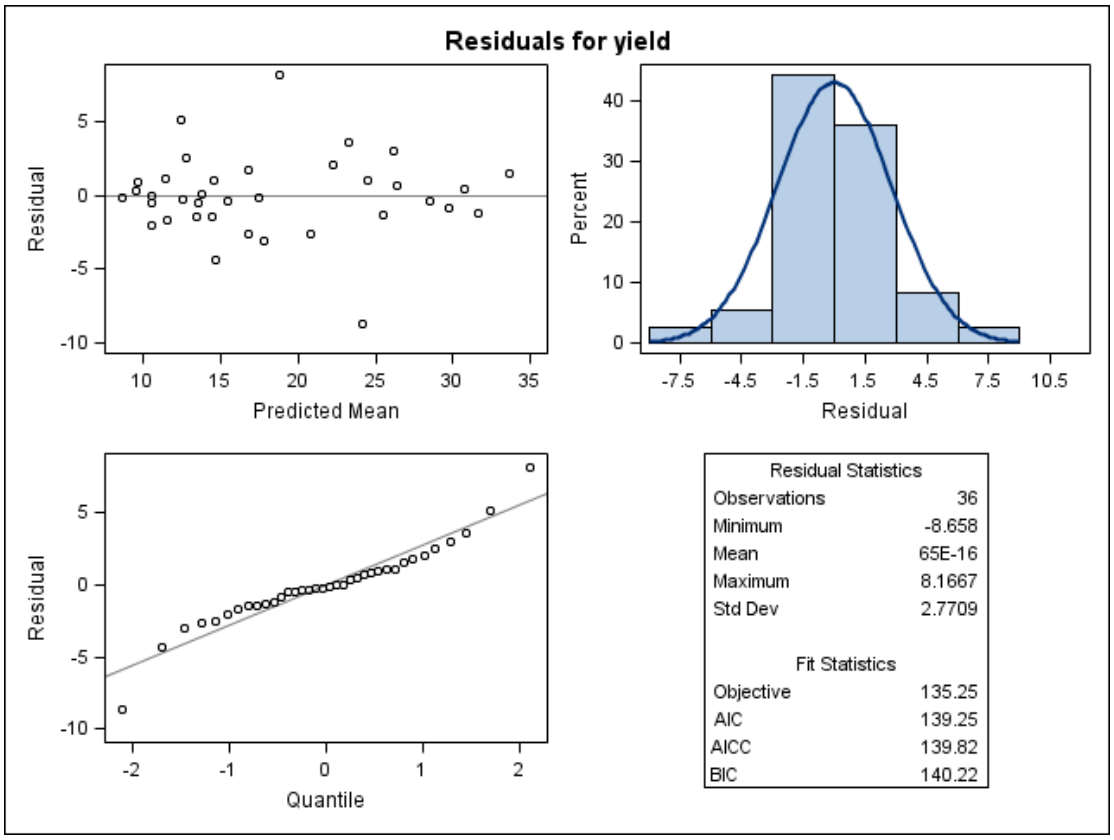
ODS GRAPHICS OFF;

The residual plots will be generated only of you use the RESIDUAL option to the MODEL statement. You can then find graphics in the Results window. An example is shown below.

```
ods graphics on;
proc mixed;

model .... / residual;

run;
ods graphics off;
```



To get the residual plots, click on the appropriate icons in the results window. For example, the "marginal residuals" $\left(y - X\hat{\beta}\right)$ are as follows:



These residuals in this example look inconspicuous.

A more direct was to obtain the residual plots is by using the option PLOTS=STUDENTPANEL in the PROC MIXED line:

```
ods graphics on;
proc mixed plots=studentpanel;

model .... / residual;

run;
ods graphics off;
```

9. PARMS statement

You can use the PARMS statement to specify starting values of all variance parameters of a mixed model. This may be very useful, e.g., if the default starting values (1 for residual and 0 for all other variance parameters) does not lead to convergence. The usage for an example with three variance components to be initialized at values 1, 5 and 10 is as follows:

PARMS (1)(5)(10);

The order in which variance components need to be specified is determined by the order of appearance of random terms in the RANDOM statement(s) of MIXED. The residual is always the last term.

You can also force MIXED to fix some or all variance components at the starting values. For example, if in the present case we want to fix the first and the third variance component at the starting values, the code is:

PARMS (1)(5)(10)/HOLD=1,3;

10. Trouble shooting in case of non-convergence

Non-convergence can have many causes. Some possible reasons are:

- The model is not correctly specified
- The covariance structure is too complex to be estimated for your data
- The default number of iterations does not suffice
- The covariance matrix estimates becomes nearly singular during iterations

The maximum number of iterations is set to 50 by default. You can increase this using the MAXITER= option to the MIXED statement. To increase the maximum number of iterations to 100, start MIXED as follows:

PROC MIXED MAXITER=100;

But do not increase to a too large number, unless you are prepared to wait for very long, only to find out that still no convergence was achieved with the maximal number of iterations.

With some covariance models, near-singularity may occur during iterations, in which case the program stops with a message in the log saying “Stopped because of infinite likelihood”. In particular, this may happen with some models for repeated measures (e.g. CSH). MIXED numerically checks in each iteration that the current estimate of the variance-covariance matrix is positive definite (i.e. not singular). The numerical check appears to be based on the determinant or some related quantity computed from the current variance-covariance matrix. If this falls below some tiny value, MIXED concludes that the matrix is singular. You can modify the threshold to achieve convergence in cases where no singularity should occur (this is the case, e.g., in CS and CSH models). This is done using the SINGULAR= option to the model statement, for example:

MODEL ….. / SINGULAR=1.E-11;

This option sometimes helps to get repeated measures variance-covariance structures to converge which do not otherwise converge.

Another option that often leads to convergence is to modify the starting values of all variance-covariance parameters using the PARMS statement (see B.9).

More hints to troubleshooting in case of numerical problems are given on the online documentation of MIXED under DETAILS -> COMPUTATIONAL ISSUES.

C. Solutions to some exercises

Exercise 2

```
data C1;
input block trt$ number_of_eggs;
datalines;
1 o   330
1 e   372
1 f   359
2 o   288
2 e   340
2 f   337
3 o   295
3 e   343
3 f   373
4 o   313
4 e   341
4 f   302
;
proc glm;
class block trt;
model number_of_eggs=block trt;
means trt/lsd;
run;

proc mixed;
class block trt;
model number_of_eggs=block trt;
lsmeans trt/pdiff;
run;

/*blocks random*/
proc glm;
class block trt;
model number_of_eggs=block trt;
random block;
means trt/lsd;
run;

proc mixed;
class block trt;
model number_of_eggs= trt;
random block;
lsmeans trt/pdiff;
run;
```

Notice that a dollar sign ($) had to be added after the variable name “trt”, because this variable is character-valued. Results are identical in each case. The mean comparison by GLM yields:

```
t Grouping          Mean      N    trt

         A        349.00      4    e
         A
         A        342.75      4    f

         B        306.50      4    o
```

Treatment O is different from E and F, but E and F are not significantly different by the LSD test.

Now I am deleting the last observation and use GLM with fixed block effects to compare means. I am comparing the MEANS and LSMEANS statements.

```
/*Data made unbalanced by deleting last obs*/
data C1;
input block trt$ number_of_eggs;
datalines;
1 o   330
1 e   372
1 f   359
2 o   288
2 e   340
2 f   337
3 o   295
3 e   343
3 f   373
4 o   313
4 e   341
4 f   .
;
/*blocks fixed*/
proc glm;
class block trt;
model number_of_eggs=block trt;
means trt/lsd;
lsmeans trt/pdiff;
run;
```

The MEANS statement yields this output:

```
                                The GLM Procedure

                        t Tests (LSD) for number_of_eggs

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error
rate.


                            Alpha                          0.05
                            Error Degrees of Freedom          5
                            Error Mean Square            179.8556
                            Critical Value of t           2.57058


          Comparisons significant at the 0.05 level are indicated by ***.


                                 Difference
                                   Between      95% Confidence
                  trt Comparison     Means          Limits

                    f - e           7.333     -18.997   33.663
                    f - o          49.833      23.503   76.163  ***
                    e - f          -7.333     -33.663   18.997
                    e - o          42.500      18.123   66.877  ***
                    o - f         -49.833     -76.163  -23.503  ***
                    o - e         -42.500     -66.877  -18.123  ***
```

Notice that the statement does not even produce means now, but only the confidence intervals for mean comparisons. This is because the data are now unbalanced and no common LSD is available.

The LSMEANS statement yields:

```
                          The GLM Procedure
                        Least Squares Means

                            number_of_      LSMEAN
                trt      eggs LSMEAN      Number

                e         349.000000           1
                f         356.083333           2
                o         306.500000           3


                Least Squares Means for effect trt
               Pr > |t| for H0: LSMean(i)=LSMean(j)

                Dependent Variable: number_of_eggs

          i/j             1             2             3

            1                        0.5337        0.0065
            2        0.5337                        0.0054
            3        0.0065        0.0054

NOTE: To ensure overall protection level, only probabilities associated with pre-planned
      comparisons should be used.
```

If you compute differences from the adjusted means, you will notice that these are not the same as those obtained from the MEANS statement, which computes simple means and their differences. Simple means are misleading in case of unbalanced data. LSMEANS is to be preferred. When data are balanced, both MEANS and LSMEANS yield the same result.

GLM does not produce estimated differences, but MIXED does:

```
proc mixed;
class block trt;
model number_of_eggs=block trt;
lsmeans trt/pdiff;
run;
```

```
                    Differences of Least Squares Means

                                      Standard
Effect    trt    _trt    Estimate       Error      DF    t Value    Pr > |t|

trt       e      f        -7.0833     10.6023       5      -0.67      0.5337
trt       e      o        42.5000      9.4830       5       4.48      0.0065
trt       f      o        49.5833     10.6023       5       4.68      0.0054
```

The p-values for the differences are the same as those for GLM, of course.

Exercise 3

```
data A2;
input
Treatmnt    Group    Block    Yield;
datalines;
    1          1          1          6
    2          1          1          7
    3          1          1          5
    4          1          1          8
    5          1          1          6
    6          1          2         16
    7          1          2         12
    8          1          2         12
    9          1          2         13
   10          1          2          8
   11          1          3         17
   12          1          3          7
   13          1          3          7
   14          1          3          9
   15          1          3         14
   16          1          4         18
   17          1          4         16
   18          1          4         13
   19          1          4         13
   20          1          4         14
   21          1          5         14
   22          1          5         15
   23          1          5         11
   24          1          5         14
   25          1          5         14
    1          2          1         24
    6          2          1         13
   11          2          1         24
   16          2          1         11
   21          2          1          8
    2          2          2         21
    7          2          2         11
   12          2          2         14
   17          2          2         11
   22          2          2         23
    3          2          3         16
    8          2          3          4
   13          2          3         12
   18          2          3         12
   23          2          3         12
    4          2          4         17
    9          2          4         10
   14          2          4         30
   19          2          4          9
   24          2          4         23
    5          2          5         15
   10          2          5         15
   15          2          5         22
   20          2          5         16
   25          2          5         19
;

/*intra-block analysis*/
proc mixed data=a2;
ods output diffs=diffs1;
class group block Treatmnt;
```

```
model Yield=group group*block Treatmnt;
lsmeans treatmnt/pdiff;
run;

data diffs1;
set diffs1;
vd=StdErr**2;

proc means data=diffs1 mean;
var vd;
run;
```

The MEANS Procedure

Analysis Variable : vd

Mean
18.2066667

```
/*with recovery of inter-block information*/
proc mixed data=a2;
ods output diffs=diffs2;
class group block Treatmnt;
model Yield=group Treatmnt/ddfm=KR;
random group*block;
lsmeans treatmnt/pdiff;
run;

data diffs2;
set diffs2;
vd=StdErr**2;

proc means data=diffs2 mean;
var vd;
run;
```

The MEANS Procedure

Analysis Variable : vd

Mean
17.9589662

Recovery of information is worthwhile because mean variance of a difference (vd) is smaller. Thus, final analysis is produced using model with random block effects. Note that the Kenward-Roger method (option ddfm=KR) is used to account for errors in the weights (estimated variance components).

Covariance Parameter Estimates

Cov Parm	Estimate
Group*Block	19.6300
Residual	13.6550

Fit Statistics

-2 Res Log Likelihood	162.9
AIC (smaller is better)	166.9
AICC (smaller is better)	167.5
BIC (smaller is better)	167.5

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Group	1	6.32	1.90	0.2152
Treatmnt	24	16.8	1.88	0.0925

Least Squares Means

Effect	Treatmnt	Estimate	Standard Error	DF	t Value	Pr > |t|
Treatmnt	1	19.0681	3.2949	23.7	5.79	<.0001
Treatmnt	2	16.9728	3.2949	23.7	5.15	<.0001
Treatmnt	3	14.6463	3.2949	23.7	4.45	0.0002
Treatmnt	4	14.7687	3.2949	23.7	4.48	0.0002
Treatmnt	5	12.8470	3.2949	23.7	3.90	0.0007
Treatmnt	6	13.1701	3.2949	23.7	4.00	0.0005
Treatmnt	7	9.0748	3.2949	23.7	2.75	0.0111
Treatmnt	8	6.7483	3.2949	23.7	2.05	0.0518
Treatmnt	9	8.3707	3.2949	23.7	2.54	0.0180
Treatmnt	10	8.4489	3.2949	23.7	2.56	0.0171
Treatmnt	11	23.5511	3.2949	23.7	7.15	<.0001
Treatmnt	12	12.4558	3.2949	23.7	3.78	0.0009
Treatmnt	13	12.6293	3.2949	23.7	3.83	0.0008
Treatmnt	14	20.7517	3.2949	23.7	6.30	<.0001
Treatmnt	15	19.3299	3.2949	23.7	5.87	<.0001
Treatmnt	16	12.6224	3.2949	23.7	3.83	0.0008
Treatmnt	17	10.5272	3.2949	23.7	3.19	0.0039
Treatmnt	18	10.7007	3.2949	23.7	3.25	0.0035
Treatmnt	19	7.3231	3.2949	23.7	2.22	0.0360
Treatmnt	20	11.4013	3.2949	23.7	3.46	0.0021
Treatmnt	21	11.6259	3.2949	23.7	3.53	0.0017
Treatmnt	22	18.5306	3.2949	23.7	5.62	<.0001
Treatmnt	23	12.2041	3.2949	23.7	3.70	0.0011

Least Squares Means

Effect	Treatmnt	Estimate	Standard Error	DF	t Value	Pr > |t|
Treatmnt	24	17.3265	3.2949	23.7	5.26	<.0001
Treatmnt	25	15.4048	3.2949	23.7	4.68	<.0001

Differences of Least Squares Means

Effect	Treatmnt	_Treatmnt	Estimate	Standard Error	DF	t Value	Pr > |t|
Treatmnt	1	2	2.0952	4.0296	17.2	0.52	0.6097
Treatmnt	1	3	4.4218	4.0296	17.2	1.10	0.2876
Treatmnt	1	4	4.2993	4.0296	17.2	1.07	0.3007
Treatmnt	1	5	6.2211	4.0296	17.2	1.54	0.1408
Treatmnt	1	6	5.8980	4.0296	17.2	1.46	0.1613
Treatmnt	1	7	9.9933	4.3382	18.2	2.30	0.0333

Etc.

There are only two standard errors of a difference, one for pairs of treatments with one direct comparison (s.e.d.=4.0296) and the other for pairs with no direct comparisons (s.e.d.=4.3382).

Exercise 4

```
data a3;
input block    response  treat;
datalines;
  1       2.4        15
  1       2.5         9
  1       2.6         1
  1       2.0        13
  2       2.7         5
  2       2.8         7
  2       2.4         8
  2       2.7         1
  3       2.6        10
  3       2.8         1
  3       2.4        14
  3       2.4         2
  4       3.4        15
  4       3.1        11
  4       2.1         2
  4       2.3         3
  5       4.1         6
  5       3.3        15
  5       3.3         4
  5       2.9         7
  6       3.4        12
  6       3.2         4
  6       2.8         3
  6       3.0         1
  7       3.2        12
  7       2.5        14
  7       2.4        15
  7       2.6         8
  8       2.3         6
  8       2.3         3
  8       2.4        14
  8       2.7         5
  9       2.8         5
  9       2.8         4
  9       2.6         2
  9       2.5        13
 10       2.5        10
 10       2.7        12
 10       2.8        13
 10       2.6         6
 11       2.6         9
```

```
 11        2.6          7
 11        2.3         10
 11        2.4          3
 12        2.7          8
 12        2.7          6
 12        2.5          2
 12        2.6          9
 13        3.0          5
 13        3.6          9
 13        3.2         11
 13        3.2         12
 14        3.0          7
 14        2.8         13
 14        2.4         14
 14        2.5         11
 15        2.4         10
 15        2.5          4
 15        3.2          8
 15        3.1         11
;
/*intra-block analysis*/
proc mixed data=a3;
ods output diffs=diffs1;
class block treat;
model response=block treat;
lsmeans treat/pdiff;
run;

data diffs1;
set diffs1;
vd=StdErr**2;

proc means data=diffs1 mean;
var vd;
run;
```

The MEANS Procedure

Analysis Variable : vd

Mean
0.0541540

```
/*with recovery of inter-block information*/
proc mixed data=a3;
ods output diffs=diffs2;
class block treat;
model response=treat/ddfm=KR;
random block;
lsmeans treat/pdiff;
run;

data diffs2;
set diffs2;
vd=StdErr**2;

proc means data=diffs2 mean;
var vd;
run;
```

The MEANS Procedure

Analysis Variable : vd

Mean
0.0512218

Recovery of information is worthwhile, so final analysis is done with random blocks model.

Cov Parm	Estimate
block	0.04652
Residual	0.08556

Fit Statistics

-2 Res Log Likelihood	52.0
AIC (smaller is better)	56.0
AICC (smaller is better)	56.3
BIC (smaller is better)	57.4

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	14	36.2	1.48	0.1676

Least Squares Means

Effect	treat	Estimate	Standard Error	DF	t Value	Pr > |t|
treat	1	2.8175	0.1686	44.1	16.71	<.0001
treat	2	2.4053	0.1686	44.1	14.27	<.0001
treat	3	2.4549	0.1686	44.1	14.56	<.0001
treat	4	2.7838	0.1686	44.1	16.51	<.0001
treat	5	2.8049	0.1686	44.1	16.63	<.0001
treat	6	2.9107	0.1686	44.1	17.26	<.0001
treat	7	2.7890	0.1686	44.1	16.54	<.0001
treat	8	2.7816	0.1686	44.1	16.50	<.0001
treat	9	2.8913	0.1686	44.1	17.15	<.0001
treat	10	2.4911	0.1686	44.1	14.77	<.0001
treat	11	2.8987	0.1686	44.1	17.19	<.0001
treat	12	3.0528	0.1686	44.1	18.11	<.0001
treat	13	2.6178	0.1686	44.1	15.53	<.0001
treat	14	2.4913	0.1686	44.1	14.78	<.0001
treat	15	2.8592	0.1686	44.1	16.96	<.0001

Differences of Least Squares Means

Effect	treat	_treat	Estimate	Standard Error	DF	t Value	Pr > |t|
treat	1	2	0.4122	0.2254	36.3	1.83	0.0757
treat	1	3	0.3626	0.2254	36.3	1.61	0.1164
treat	1	4	0.03369	0.2254	36.3	0.15	0.8820
treat	1	5	0.01262	0.2254	36.3	0.06	0.9556
treat	1	6	-0.09317	0.2317	37.7	-0.40	0.6898

treat	1	7	0.02854	0.2254	36.3	0.13	0.9000
treat	1	8	0.03592	0.2254	36.3	0.16	0.8743
treat	1	9	-0.07379	0.2254	36.3	-0.33	0.7453
treat	1	10	0.3265	0.2254	36.3	1.45	0.1561
treat	1	11	-0.08118	0.2317	37.7	-0.35	0.7280
treat	1	12	-0.2353	0.2254	36.3	-1.04	0.3035
treat	1	13	0.1998	0.2254	36.3	0.89	0.3814
treat	1	14	0.3262	0.2254	36.3	1.45	0.1565
treat	1	15	-0.04171	0.2254	36.3	-0.19	0.8542
treat	2	3	-0.04963	0.2254	36.3	-0.22	0.8270
treat	2	4	-0.3785	0.2254	36.3	-1.68	0.1017
treat	2	5	-0.3996	0.2254	36.3	-1.77	0.0847
treat	2	6	-0.5054	0.2254	36.3	-2.24	0.0312
treat	2	7	-0.3837	0.2317	37.7	-1.66	0.1060
treat	2	8	-0.3763	0.2254	36.3	-1.67	0.1037
treat	2	9	-0.4860	0.2254	36.3	-2.16	0.0378
treat	2	10	-0.08575	0.2254	36.3	-0.38	0.7059
treat	2	11	-0.4934	0.2254	36.3	-2.19	0.0351
treat	2	12	-0.6475	0.2317	37.7	-2.80	0.0081
treat	2	13	-0.2125	0.2254	36.3	-0.94	0.3522
treat	2	14	-0.08600	0.2254	36.3	-0.38	0.7051
treat	2	15	-0.4539	0.2254	36.3	-2.01	0.0515
treat	3	4	-0.3289	0.2254	36.3	-1.46	0.1532
treat	3	5	-0.3500	0.2254	36.3	-1.55	0.1292
treat	3	6	-0.4558	0.2254	36.3	-2.02	0.0506
treat	3	7	-0.3340	0.2254	36.3	-1.48	0.1470
treat	3	8	-0.3267	0.2317	37.7	-1.41	0.1667
treat	3	9	-0.4364	0.2254	36.3	-1.94	0.0607
treat	3	10	-0.03612	0.2254	36.3	-0.16	0.8736
treat	3	11	-0.4438	0.2254	36.3	-1.97	0.0567
treat	3	12	-0.5979	0.2254	36.3	-2.65	0.0118
treat	3	13	-0.1628	0.2317	37.7	-0.70	0.4865
treat	3	14	-0.03637	0.2254	36.3	-0.16	0.8727
treat	3	15	-0.4043	0.2254	36.3	-1.79	0.0812
treat	4	5	-0.02107	0.2254	36.3	-0.09	0.9261
treat	4	6	-0.1269	0.2254	36.3	-0.56	0.5770
treat	4	7	-0.00515	0.2254	36.3	-0.02	0.9819
treat	4	8	0.002225	0.2254	36.3	0.01	0.9922
treat	4	9	-0.1075	0.2317	37.7	-0.46	0.6453
treat	4	10	0.2928	0.2254	36.3	1.30	0.2022
treat	4	11	-0.1149	0.2254	36.3	-0.51	0.6134
treat	4	12	-0.2690	0.2254	36.3	-1.19	0.2405
treat	4	13	0.1661	0.2254	36.3	0.74	0.4661
treat	4	14	0.2925	0.2317	37.7	1.26	0.2145
treat	4	15	-0.07540	0.2254	36.3	-0.33	0.7399
treat	5	6	-0.1058	0.2254	36.3	-0.47	0.6416
treat	5	7	0.01591	0.2254	36.3	0.07	0.9441
treat	5	8	0.02329	0.2254	36.3	0.10	0.9183
treat	5	9	-0.08641	0.2254	36.3	-0.38	0.7037
treat	5	10	0.3138	0.2317	37.7	1.35	0.1836
treat	5	11	-0.09380	0.2254	36.3	-0.42	0.6798
treat	5	12	-0.2479	0.2254	36.3	-1.10	0.2787
treat	5	13	0.1871	0.2254	36.3	0.83	0.4119
treat	5	14	0.3136	0.2254	36.3	1.39	0.1727
treat	5	15	-0.05434	0.2317	37.7	-0.23	0.8158
treat	6	7	0.1217	0.2254	36.3	0.54	0.5926
treat	6	8	0.1291	0.2254	36.3	0.57	0.5704
treat	6	9	0.01938	0.2254	36.3	0.09	0.9320
treat	6	10	0.4196	0.2254	36.3	1.86	0.0708
treat	6	11	0.01199	0.2317	37.7	0.05	0.9590
treat	6	12	-0.1421	0.2254	36.3	-0.63	0.5323
treat	6	13	0.2929	0.2254	36.3	1.30	0.2020
treat	6	14	0.4194	0.2254	36.3	1.86	0.0709

treat	6	15	0.05146	0.2254	36.3	0.23	0.8207
treat	7	8	0.007380	0.2254	36.3	0.03	0.9741
treat	7	9	-0.1023	0.2254	36.3	-0.45	0.6526
treat	7	10	0.2979	0.2254	36.3	1.32	0.1946
treat	7	11	-0.1097	0.2254	36.3	-0.49	0.6294
treat	7	12	-0.2638	0.2317	37.7	-1.14	0.2619
treat	7	13	0.1712	0.2254	36.3	0.76	0.4524
treat	7	14	0.2977	0.2254	36.3	1.32	0.1949
treat	7	15	-0.07025	0.2254	36.3	-0.31	0.7571
treat	8	9	-0.1097	0.2254	36.3	-0.49	0.6294
treat	8	10	0.2905	0.2254	36.3	1.29	0.2056
treat	8	11	-0.1171	0.2254	36.3	-0.52	0.6066
treat	8	12	-0.2712	0.2254	36.3	-1.20	0.2367
treat	8	13	0.1638	0.2317	37.7	0.71	0.4838
treat	8	14	0.2903	0.2254	36.3	1.29	0.2060
treat	8	15	-0.07763	0.2254	36.3	-0.34	0.7326
treat	9	10	0.4002	0.2254	36.3	1.78	0.0842
treat	9	11	-0.00739	0.2254	36.3	-0.03	0.9740
treat	9	12	-0.1615	0.2254	36.3	-0.72	0.4783
treat	9	13	0.2735	0.2254	36.3	1.21	0.2328
treat	9	14	0.4000	0.2317	37.7	1.73	0.0924
treat	9	15	0.03208	0.2254	36.3	0.14	0.8876
treat	10	11	-0.4076	0.2254	36.3	-1.81	0.0789
treat	10	12	-0.5618	0.2254	36.3	-2.49	0.0174
treat	10	13	-0.1267	0.2254	36.3	-0.56	0.5775
treat	10	14	-0.00025	0.2254	36.3	-0.00	0.9991
treat	10	15	-0.3682	0.2317	37.7	-1.59	0.1204
treat	11	12	-0.1541	0.2254	36.3	-0.68	0.4985
treat	11	13	0.2809	0.2254	36.3	1.25	0.2207
treat	11	14	0.4074	0.2254	36.3	1.81	0.0790
treat	11	15	0.03947	0.2254	36.3	0.18	0.8620
treat	12	13	0.4351	0.2254	36.3	1.93	0.0615
treat	12	14	0.5615	0.2254	36.3	2.49	0.0175
treat	12	15	0.1936	0.2254	36.3	0.86	0.3961
treat	13	14	0.1265	0.2254	36.3	0.56	0.5783
treat	13	15	-0.2415	0.2254	36.3	-1.07	0.2912
treat	14	15	-0.3679	0.2254	36.3	-1.63	0.1113

To get a compact letter display for all pairwise comparisons, we use the GLIMMIX procedure as follows:

```
proc glimmix data=a3;
class block treat;
model response=treat/ddfm=KR;
random block;
lsmeans treat/pdiff lines;
run;
```

T Grouping for treat Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

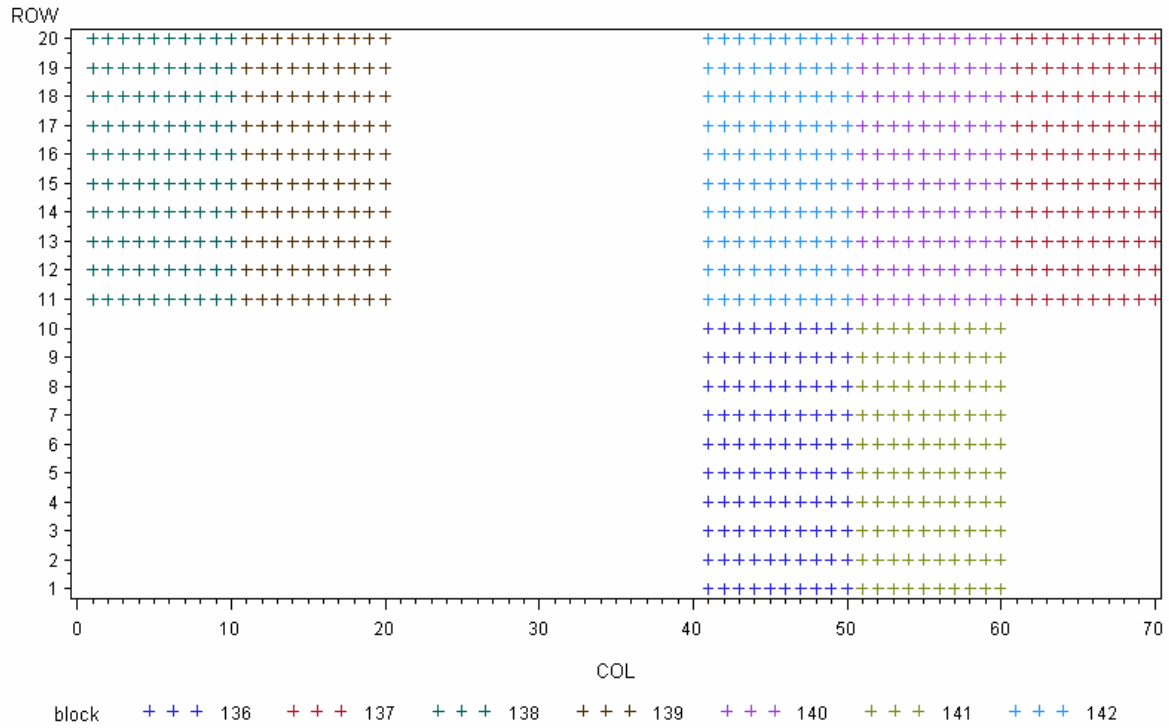
treat	Estimate		
12	3.0528		A
			A
6	2.9107	B	A
		B	A
11	2.8987	B	A
		B	A
9	2.8913	B	A
		B	A

15	2.8592	B	A	C
		B	A	C
1	2.8175	B	A	C
		B	A	C
5	2.8049	B	A	C
		B	A	C
7	2.7890	B	A	C
		B	A	C
4	2.7838	B	A	C
		B	A	C
8	2.7816	B	A	C
		B	A	C
13	2.6178	B	A	C
		B		C
14	2.4913	B		C
		B		C
10	2.4911	B		C
		B		C
3	2.4549	B		C
				C
2	2.4053			C

Note that the F-test for treatments is not significant (p=0.1676), so the few significant differences obtained by t-tests must be taken with caution (they are likely to be false positives, i.e. false rejections of the null hypothesis).

Exercise 5

```
proc gplot data=a4;
plot row*col=block;
run;
```



There are only 7 blocks, so recovery of information may not be worthwhile. As regards effects for rows and columns (post blocking) there are two options: (i) fit effects for roes and columns across the whole experimental area; (ii) nest row and column effects within blocks.

The second option is reasonable, because blocks probably have been managed consecutively, i.e. one after the other.

Model	Mean v.d.	AIC[&]
Blocks fixed	4697.64	-
Blocks fixed, row+col fixed[&]		-
Blocks random	4678.20	1206.6
Blocks random, block*row+block*col random[$]	-	1208.3
Blocks random, row+col random[%]	-	1206.7

[&] Only can compare models with same fixed effects
[$] row and column effects nested within blocks
[%] row and column effects fitted across whole field

Post-blocking with row and column effects does not seem worthwhile here, because the AIC does not decrease, so the usual incomplete block analysis is preferred.

Important note on computing time: Computation of F-statistics as well as of adjusted means is very time consuming with this dataset. When comparing just AIC values of different models, these things are not needed, so they may be switched off. For the random blocks model the code with the appropriate options is as follows:

```
proc mixed data=a4 lognote;
class block cultivar;
model yield= cultivar / ddfm=residual notest;
random block;
run;
```

The NOTEST option switches off the F-tests. The DDFM=residual selects a much simpler method for computing degrees of freedom. The LOGNOTE option prints convergence status messages to the LOG-window, which is convenient for monitoring computational progress.

The full code for the same model that also computes adjusted means and average variance of a difference, but takes very long to compute (about 10 minutes or more!), is as follows:

```
/*analysis of incomplete block design with random blocks*/
proc mixed data=a4 lognote;
ods output diffs=diffs2;
class block cultivar;
model yield= cultivar / ddfm=kr;
random block;
lsmeans cultivar/ pdiff;
run;

data diffs2;
set diffs2;
vd=StdErr**2;
```

```
proc means data=diffs2 mean;
var vd;
run;
```

Analysis with random blocks:

```
Covariance Parameter
      Estimates

Cov Parm     Estimate

block          670.44
Residual      2263.17
```

Block variance is small when blocks are taken as random, which is why recovery of information is worthwhile despite the small number of blocks.

Exercise 6

A model with fixed blocks gives a larger standard error of a difference (s.e.d.= 3.5024) than a model with random blocks (s.e.d.= 3.4157). In both cases, the s.e.d. is constant. In this sense, the design is balanced, i.e., each mean difference has the same variance or standard error of a difference. This property of a design is also called **variance-balanced**.

```
data a14;
input
block hybrid yield;
datalines;
1 3  25.3
1 6  19.9
1 9  29.0
1 11 24.6
2 3  23.0
2 4  19.8
2 8  33.3
2 12 22.7
3 10 16.2
3 11 19.3
3 12 31.7
3 13 26.6
4 2 27.3
4 5 27.0
4 8 35.6
4 11 17.4
5 7 23.4
5 8 30.5
5 9 30.8
5 10 32.4
6 4 30.6
6 5 32.4
6 6 27.2
6 10 32.8
7 1 34.7
7 5 31.1
7 9 25.7
7 12 30.5
8 3 34.4
8 5 32.4
8 7 33.3
```

```
8 13 36.9
9 1 38.2
9 2 32.9
9 3 37.3
9 10 31.3
10 2 28.7
10 4  30.7
10 9 26.9
10 13 35.3
11 1 36.6
11 4 31.1
11 7 31.1
11 11 28.4
12 1 31.8
12 6 33.7
12 8 27.8
12 13 41.1
13 2 30.3
13 6 31.5
13 7 39.3
13 12 26.7
;
/*intra-block analysis*/
proc mixed data=a14;
class block hybrid;
model yield=block hybrid/ddfm=kr;
lsmeans hybrid/pdiff;
run;

/*with inter-block analysis*/
proc mixed data=a14;
class block hybrid;
model yield=hybrid/ddfm=kr;
random block;
lsmeans hybrid/pdiff;
run;
```

Exercise 7

```
data c2;
input block trt pulserate;
datalines;
1 12 75
1 13 87
1 21 84
1 22 93
1 23 99
2 11 93
2 12 84
2 13 96
2 21 90
2 22 108
3 11 99
3 12 93
3 13 96
3 22 123
3 23 129
4 11 99
4 12 108
4 13 99
4 21 99
4 23 120
```

```
5 11 99
5 13 111
5 21 90
5 22 129
5 23 141
6 11 129
6 12 135
6 21 120
6 22 147
6 23 153
;

/*intra-block analysis - blocks fixed*/
proc mixed data=c2;
class block trt;
model pulserate=trt block;
lsmeans trt/pdiff;
run;

/*with recovery of inter-block information - blocks random*/
proc mixed data=c2;
class block trt;
model pulserate=trt;
random block;
lsmeans trt/priff;
run;
```

Results of mean comparisons with recovery of information:

Differences of Least Squares Means

Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > |t|
trt	11	12	-0.3250	4.2865	19	-0.08	0.9404
trt	11	13	-3.6697	4.2865	19	-0.86	0.4026
trt	11	21	3.3805	4.2865	19	0.79	0.4401
trt	11	22	-20.3580	4.2865	19	-4.75	0.0001
trt	11	23	-26.9207	4.2865	19	-6.28	<.0001
trt	12	13	-3.3448	4.2865	19	-0.78	0.4448
trt	12	21	3.7054	4.2865	19	0.86	0.3981
trt	12	22	-20.0330	4.2865	19	-4.67	0.0002
trt	12	23	-26.5958	4.2865	19	-6.20	<.0001
trt	13	21	7.0502	4.2865	19	1.64	0.1165
trt	13	22	-16.6882	4.2865	19	-3.89	0.0010
trt	13	23	-23.2510	4.2865	19	-5.42	<.0001
trt	21	22	-23.7384	4.2865	19	-5.54	<.0001
trt	21	23	-30.3012	4.2865	19	-7.07	<.0001
trt	22	23	-6.5627	4.2865	19	-1.53	0.1422

The standard error of a difference (s.e.d.) is constant, meaning that the design is variance-balanced. This is why the design is called “balanced incomplete block design”.

Tha analysis so far has ignored the factorial structure of the experiment. For a two-factorial analysis, we need to code the treatment factors C and D explicitly.

```
data c2;
input block trt C D pulserate;
datalines;
1 12 1 2 75
1 13 1 3 87
```

```
1 21 2 1 84
1 22 2 2 93
1 23 2 3 99
2 11 1 1 93
2 12 1 2 84
2 13 1 3 96
2 21 2 1 90
2 22 2 2 108
3 11 1 1 99
3 12 1 2 93
3 13 1 3 96
3 22 2 2 123
3 23 2 3 129
4 11 1 1 99
4 12 1 2 108
4 13 1 3 99
4 21 2 1 99
4 23 2 3 120
5 11 1 1 99
5 13 1 3 111
5 21 2 1 90
5 22 2 2 129
5 23 2 3 141
6 11 1 1 129
6 12 1 2 135
6 21 2 1 120
6 22 2 2 147
6 23 2 3 153
;
/*with recovery of inter-block information - blocks random*/
proc mixed data=c2;
class block C D;
model pulserate=C D C*D;
random block;
run;
```

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
C	1	19	27.05	<.0001
D	2	19	15.22	0.0001
C*D	2	19	11.18	0.0006

Interaction is significant, so we need to compare C*D means.

```
/*compare C*D means because interaction was significant*/
proc mixed data=c2;
class block C D;
model pulserate=C D C*D;
random block;
lsmeans C*D/pdiff;
run;
```

Differences of Least Squares Means

Effect	C	D	_C	_D	Estimate	Standard Error	DF	t Value	Pr > |t|
C*D	1	1	1	2	-0.3250	4.2865	19	-0.08	0.9404
C*D	1	1	1	3	-3.6697	4.2865	19	-0.86	0.4026
C*D	1	1	2	1	3.3805	4.2865	19	0.79	0.4401
C*D	1	1	2	2	-20.3580	4.2865	19	-4.75	0.0001
C*D	1	1	2	3	-26.9207	4.2865	19	-6.28	<.0001
C*D	1	2	1	3	-3.3448	4.2865	19	-0.78	0.4448
C*D	1	2	2	1	3.7054	4.2865	19	0.86	0.3981
C*D	1	2	2	2	-20.0330	4.2865	19	-4.67	0.0002
C*D	1	2	2	3	-26.5958	4.2865	19	-6.20	<.0001
C*D	1	3	2	1	7.0502	4.2865	19	1.64	0.1165
C*D	1	3	2	2	-16.6882	4.2865	19	-3.89	0.0010
C*D	1	3	2	3	-23.2510	4.2865	19	-5.42	<.0001
C*D	2	1	2	2	-23.7384	4.2865	19	-5.54	<.0001
C*D	2	1	2	3	-30.3012	4.2865	19	-7.07	<.0001
C*D	2	2	2	3	-6.5627	4.2865	19	-1.53	0.1422

You can use the ODS system to filter out comparisons at the same levels of C.

```
/*filter out comparisons at same level of C*/
proc mixed data=c2;
ods output diffs=diffs_C;
class block C D;
model pulserate=C D C*D;
random block;
lsmeans C*D/pdiff;
run;

data diffs_C;
set diffs_C;
if C = _C;
run;

proc print data=diffs_C;
run;
```

Obs	Effect	C	D	_C	_D	Estimate	StdErr	DF	tValue	Probt
1	C*D	1	1	1	2	-0.3250	4.2865	19	-0.08	0.9404
2	C*D	1	1	1	3	-3.6697	4.2865	19	-0.86	0.4026
3	C*D	1	2	1	3	-3.3448	4.2865	19	-0.78	0.4448
4	C*D	2	1	2	2	-23.7384	4.2865	19	-5.54	<.0001
5	C*D	2	1	2	3	-30.3012	4.2865	19	-7.07	<.0001
6	C*D	2	2	2	3	-6.5627	4.2865	19	-1.53	0.1422

The same for comparisons at the same level of D:

```
/*filter out comparisons at same level of D*/
proc mixed data=c2;
ods output diffs=diffs_D;
class block C D;
model pulserate=C D C*D;
random block;
lsmeans C*D/pdiff;
run;

data diffs_D;
set diffs_D;
if D = _D;
```

```
run;

proc print data=diffs_D;
run;
```

Obs	Effect	C	D	_C	_D	Estimate	StdErr	DF	tValue	Probt
1	C*D	1	1	2	1	3.3805	4.2865	19	0.79	0.4401
2	C*D	1	2	2	2	-20.0330	4.2865	19	-4.67	0.0002
3	C*D	1	3	2	3	-23.2510	4.2865	19	-5.42	<.0001

Exercise 8

(1) It was believed that roasts from similar positions on the two sides of the animal would be similar. This means that using animal as a block variable can improve precision.

(2) Each pair occurs exactly once. There are $\binom{6}{2} = \frac{6!}{4!2!} = 15$ different pairs of 2 out of 6 treatments, making 15 possible blocks. The design is variance-balanced, because each pair of treatments has the same number of associations (direct comparisons).

(3)

```
data C3;
input block    time     y;
datalines;
   1        1       7
   1        2      17
   2        3      26
   2        4      25
   3        5      33
   3        6      29
   4        1      17
   4        3      27
   5        2      23
   5        5      27
   6        4      29
   6        6      30
   7        1      10
   7        4      25
   8        2      26
   8        6      37
   9        3      24
   9        5      26
  10        1      25
  10        5      40
  11        2      25
  11        4      34
  12        3      34
  12        6      32
  13        1      11
  13        6      27
  14        2      24
  14        3      21
  15        4      26
  15        5      32
;
/*mean comparisons - intra-block analysis*/
proc mixed data=c3;
```

```
class block time;
model y=block time;
lsmeans time/pdiff cl;
run;
```

The option “CL to the LSMEANS statement computes 95% confidence limits, while the option PDIFF makes sure that differences are computed, plus p-values of t-tests of H0: Differences are zero.

Differences of Least Squares Means

Effect	time	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	1	2	-9.1667	2.2706	10	-4.04	0.0024	0.05	-14.2258	-4.1075
time	1	3	-12.3333	2.2706	10	-5.43	0.0003	0.05	-17.3925	-7.2742
time	1	4	-13.6667	2.2706	10	-6.02	0.0001	0.05	-18.7258	-8.6075
time	1	5	-16.1667	2.2706	10	-7.12	<.0001	0.05	-21.2258	-11.1075
time	1	6	-14.6667	2.2706	10	-6.46	<.0001	0.05	-19.7258	-9.6075
time	2	3	-3.1667	2.2706	10	-1.39	0.1933	0.05	-8.2258	1.8925
time	2	4	-4.5000	2.2706	10	-1.98	0.0756	0.05	-9.5592	0.5592
time	2	5	-7.0000	2.2706	10	-3.08	0.0116	0.05	-12.0592	-1.9408
time	2	6	-5.5000	2.2706	10	-2.42	0.0359	0.05	-10.5592	-0.4408
time	3	4	-1.3333	2.2706	10	-0.59	0.5701	0.05	-6.3925	3.7258
time	3	5	-3.8333	2.2706	10	-1.69	0.1223	0.05	-8.8925	1.2258
time	3	6	-2.3333	2.2706	10	-1.03	0.3283	0.05	-7.3925	2.7258
time	4	5	-2.5000	2.2706	10	-1.10	0.2967	0.05	-7.5592	2.5592
time	4	6	-1.0000	2.2706	10	-0.44	0.6690	0.05	-6.0592	4.0592
time	5	6	1.5000	2.2706	10	0.66	0.5238	0.05	-3.5592	6.5592

Note that the s.e.d. is constant because the design is variance-balanced.

(4)

```
/*the contrast - intra-block analysis*/
proc mixed data=c3;
class block time;
model y=block time;
estimate '4+5+6 vs 1+2+3' time -1 -1 -1 1 1 1/divisor=3;
contrast '4+5+6 vs 1+2+3' time -1 -1 -1 1 1 1;
run;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
4+5+6 vs 1+2+3	7.6667	1.3109	10	5.85	0.0002

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
4+5+6 vs 1+2+3	1	10	34.20	0.0002

Longer storage gives better tenderness scores. The same p-value of significance tests is obtained with both statements, though the ESTIMATE statement uses a t-test, while the CONTRAST statement uses an F-test. Both tests are, in fact, equivalent in this case. In the ESTIMATE statement we can use the DIVISOR option to specify the common denominator

of all contrast coefficients, while this is neither possible nor necessary with the CONTRAST statement, which does not produce the contrast estimate itself.

(5) Simple arithmetic treatment means are inappropriate because the design involves incomplete blocks. Simple means do not account for and block adjustments.

(6) & (7) The ANOVA shows a significant test for the TIME effect (p=0.0004), suggesting that there are real differences among the adjusted means. There is a significant linear increase of tenderness with time based on the linear contrast. The same analysis can be obtained with MIXED, both with and without recovery of information ($p < 0.0001$).

```
 /*mean comparisons and linear contrast - intra-block analysis*/
proc mixed data=c3;
class block time;
model y=block time;
lsmeans time/pdiff cl;
estimate 'linear TIME' time -5 -3 -1 1 3 5;
contrast 'linear TIME' time -5 -3 -1 1 3 5;
run;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
linear TIME	95.6667	13.4330	10	7.12	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
linear TIME	1	10	50.72	<.0001

Least Squares Means

Effect	time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	1	14.6333	1.5511	10	9.43	<.0001	0.05	11.1773	18.0894
time	2	23.8000	1.5511	10	15.34	<.0001	0.05	20.3439	27.2561
time	3	26.9667	1.5511	10	17.39	<.0001	0.05	23.5106	30.4227
time	4	28.3000	1.5511	10	18.25	<.0001	0.05	24.8439	31.7561
time	5	30.8000	1.5511	10	19.86	<.0001	0.05	27.3439	34.2561
time	6	29.3000	1.5511	10	18.89	<.0001	0.05	25.8439	32.7561

Differences of Least Squares Means

Effect	time	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	1	2	-9.1667	2.2706	10	-4.04	0.0024	0.05	-14.2258	-4.1075
time	1	3	-12.3333	2.2706	10	-5.43	0.0003	0.05	-17.3925	-7.2742
time	1	4	-13.6667	2.2706	10	-6.02	0.0001	0.05	-18.7258	-8.6075
time	1	5	-16.1667	2.2706	10	-7.12	<.0001	0.05	-21.2258	-11.1075
time	1	6	-14.6667	2.2706	10	-6.46	<.0001	0.05	-19.7258	-9.6075
time	2	3	-3.1667	2.2706	10	-1.39	0.1933	0.05	-8.2258	1.8925
time	2	4	-4.5000	2.2706	10	-1.98	0.0756	0.05	-9.5592	0.5592
time	2	5	-7.0000	2.2706	10	-3.08	0.0116	0.05	-12.0592	-1.9408
time	2	6	-5.5000	2.2706	10	-2.42	0.0359	0.05	-10.5592	-0.4408
time	3	4	-1.3333	2.2706	10	-0.59	0.5701	0.05	-6.3925	3.7258
time	3	5	-3.8333	2.2706	10	-1.69	0.1223	0.05	-8.8925	1.2258
time	3	6	-2.3333	2.2706	10	-1.03	0.3283	0.05	-7.3925	2.7258

time	4	5	-2.5000	2.2706	10	-1.10	0.2967	0.05	-7.5592	2.5592
time	4	6	-1.0000	2.2706	10	-0.44	0.6690	0.05	-6.0592	4.0592
time	5	6	1.5000	2.2706	10	0.66	0.5238	0.05	-3.5592	6.5592

```
/*mean comparisons and linear contrast - with recovery of inter-block
information*/
proc mixed data=c3;
class block time;
model y=time;
random block;
lsmeans time/pdiff cl;
estimate 'linear TIME' time -5 -3 -1 1 3 5;
contrast 'linear TIME' time -5 -3 -1 1 3 5;
run;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
linear TIME	100.33	12.4076	10	8.09	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
linear TIME	1	10	65.39	<.0001

Least Squares Means

Effect	time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	1	14.4547	1.7699	10	8.17	<.0001	0.05	10.5112	18.3982
time	2	23.5744	1.7699	10	13.32	<.0001	0.05	19.6309	27.5179
time	3	26.8069	1.7699	10	15.15	<.0001	0.05	22.8633	30.7504
time	4	28.1590	1.7699	10	15.91	<.0001	0.05	24.2155	32.1025
time	5	31.0256	1.7699	10	17.53	<.0001	0.05	27.0821	34.9691
time	6	29.7794	1.7699	10	16.83	<.0001	0.05	25.8359	33.7229

Differences of Least Squares Means

Effect	time	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
time	1	2	-9.1197	2.0973	10	-4.35	0.0014	0.05	-13.7927	-4.4467
time	1	3	-12.3521	2.0973	10	-5.89	0.0002	0.05	-17.0251	-7.6791
time	1	4	-13.7043	2.0973	10	-6.53	<.0001	0.05	-18.3773	-9.0313
time	1	5	-16.5709	2.0973	10	-7.90	<.0001	0.05	-21.2439	-11.8979
time	1	6	-15.3247	2.0973	10	-7.31	<.0001	0.05	-19.9977	-10.6517
time	2	3	-3.2325	2.0973	10	-1.54	0.1543	0.05	-7.9055	1.4405
time	2	4	-4.5846	2.0973	10	-2.19	0.0537	0.05	-9.2576	0.08838
time	2	5	-7.4512	2.0973	10	-3.55	0.0052	0.05	-12.1242	-2.7782
time	2	6	-6.2050	2.0973	10	-2.96	0.0143	0.05	-10.8780	-1.5321
time	3	4	-1.3521	2.0973	10	-0.64	0.5336	0.05	-6.0251	3.3209
time	3	5	-4.2188	2.0973	10	-2.01	0.0720	0.05	-8.8917	0.4542
time	3	6	-2.9726	2.0973	10	-1.42	0.1868	0.05	-7.6456	1.7004
time	4	5	-2.8666	2.0973	10	-1.37	0.2016	0.05	-7.5396	1.8064
time	4	6	-1.6204	2.0973	10	-0.77	0.4576	0.05	-6.2934	3.0526
time	5	6	1.2462	2.0973	10	0.59	0.5656	0.05	-3.4268	5.9192

(8) This requires adding a lack-of-tit variable (LOF) that is equivalent to the treatment variable TIME. TIME itself is used for the linear regression, while the LOF is fitted as a CLASS variable. The model is

$$y_{ij} = \mu + b_j + \beta t_i + \delta_i + e_{ij},$$

where b_j is the j-th block effect, β is the regression coefficient for the linear regression on time t_i (modelled by TIME), and δ_i is the lack-of-fit effect (modelled by LOF).

```
/*Lack-of-fit test*/
data c3;
set c3;
LOF=time;
run;

/*LOF test - with recovery of inter-block information*/
proc mixed data=c3;
class block LOF;
model y=time LOF;
random block;
run;
```

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	0	.	.	.
LOF	4	10	4.48	0.0248

There is a significant lack-of-fit, meaning that the linear regression does not adequately describe the data. No test is provided for TIME, because the default method uses so-called Type III tests. We can switch to sequential F-tests (Type I tests) using the option HTYPE=1.

```
proc mixed data=c3;
class block LOF;
model y=time LOF/htype=1;
random block;
run;
```

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	1	10	65.39	<.0001
LOF	4	10	4.48	0.0248

Because of the significant lack of fit, we need to modify the model. One option is to add a quadratic term, so the model becomes

$$y_{ij} = \mu + b_j + \beta t_i + \gamma t_i^2 + \delta_i + e_{ij}$$

```
/*LOF test for quadratic regression model - with recovery of inter-block
information*/
proc mixed data=c3;
class block LOF;
model y=time time*time LOF/htype=1;
random block;
run;
```

```
        Type 1 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F

time            1      10      65.39    <.0001
time*time       1      10      15.37    0.0029
LOF             3      10       0.85    0.4973
```

Now the lack of fit is not significant, so the quadratic model is appropriate. To estimate the final model, we drop the lack-of-fit term and rerun the analysis, using the SOLUTION option to print the fixed effect estimates.

```
/*Estimate final model - with recovery of inter-block information*/
proc mixed data=c3;
class block;
model y=time time*time/solution;
random block;
run;
```

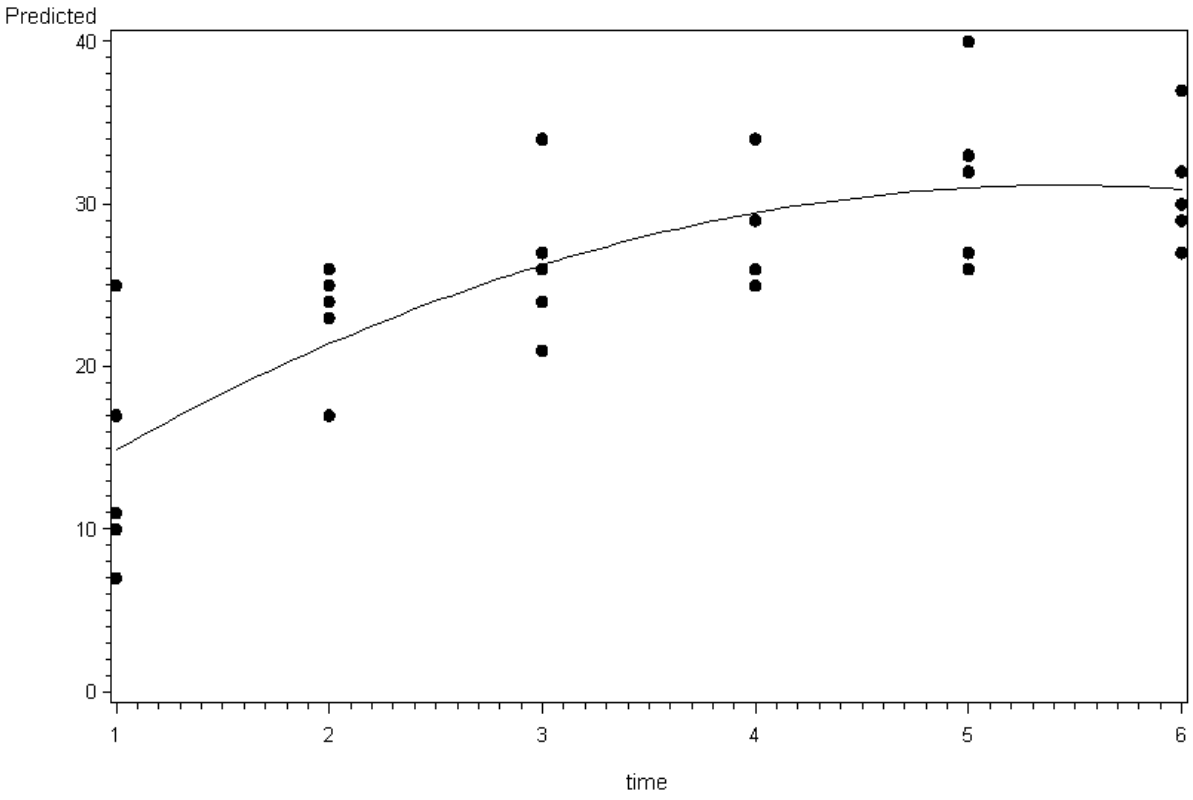
```
                  Solution for Fixed Effects

                        Standard
Effect       Estimate       Error      DF    t Value    Pr > |t|

Intercept      6.7189      2.7831      14       2.41      0.0300
time           9.5364      1.7051      13       5.59      <.0001
time*time     -0.9536      0.2384      13      -4.00      0.0015
```

The fitted model is

Tenderness = 6.7189 + 9.5364*TIME – 0.9536*TIME^2

The following code and the graphics export facility can be used to generate a plot:

```
/*Show model fit*/
proc mixed data=c3;
class block;
model y=time time*time/solution outp=pred;
random block;
run;

symbol1 i=rq value=none color=black;
symbol2 i=none value=dot color=black;
proc gplot data=pred;
plot pred*time y*time/overlay;
run;
```



Exercise 9

(i) There will be a total of 10*3 = 30 experimental unit. If each of the 6 treatments is equally important then each will need to have 5 replications.
(ii) Each treatment will occur in 5 blocks
(iii) In each block, in which a treatment x occurs, there will be 3 direct comparisons with other treatments. This means, that each treatment will have a total of 10 direct comparisons to the other five treatments
(iv) There will be two direct comparisons with each other treatment. In other words, each treatment pair occurs together in the same block twice.

Step (1): Place treatment A in first 5 blocks.

	Block									
	I	II	III	IV	V	VI	VII	VIII	IX	X
A	x	x	x	x	x	.	.	.	.	.
B	.	.	.	.	.	.	.	.	.	.
C	.	.	.	.	.	.	.	.	.	.
D		.	.	.	.	.	.	.	.	.
E	.	.	.	.	.	.	.	.	.	.
F	.	.	.	.	.	.	.	.	.	.

Step (2) fill up first five blocks so that each block has size three and there are two direct comparisons of A with each other treatment.

Block

	I	II	III	IV	V	VI	VII	VIII	IX	X
A	x	x	x	x	x	.	.	.	.	.
B	x	x	.	.	.	.	.	.	.	.
C	x	.	x	.	.	.	.	.	.	.
D		x	.	x	.	.	.	.	.	.
E	.	.	x	.	x	.	.	.	.	.
F	.	.	.	x	x	.	.	.	.	.

Step (3): Fill up the remaining block, making sure the number of direct comparisons is two for each pair.

Block

	I	II	III	IV	V	VI	VII	VIII	IX	X
A	x	x	x	x	x	.	.	.	.	.
B	x	x		.	.	x	x	x	.	.
C	x	.	x	.	.	x	.	.	x	x
D	.	x		x	.	.	.	.	.	.
E	.	.	x	.	x	.	.	.	.	.
F	.	.	.	x	x	.	.	.	.	.

Block

	I	II	III	IV	V	VI	VII	VIII	IX	X
A	x	x	x	x	x	.	.	.	.	.
B	x	x	.	.	.	x	x	x	.	.
C	x	.	x	.	.	x	.	.	x	x
D	.	x		x	.	.	x	.	x	x
E	.	.	x	.	x	.	x	x	x	.
F	.	.	.	x	x	x	.	x	.	x

Each pair now has two direct comparisons. The standard error of a difference is constant, meaning the design is variance balanced. This can be verified by the following dummy analysis that fixed the error variance at 1.

```
data C4;
do trt=1 to 6;
  do block=1 to 10;
    input a$ @@;
    if a='x' then do;
        y=1; output;
      end;
  end;
end;
datalines;
x     x     x     x     x     .     .     .     .     .
x     x     .     .     .     x     x     x     .     .
x     .     x     .     .     x     .     .     x     x
.     x     .     x     .     .     x     .     x     x
.     .     x     .     x     .     x     x     x     .
.     .     .     x     x     x     .     x     .     x
;
```

```
proc mixed data=c4;
class trt block;
model y=trt block;
parms (1)/hold=1;
lsmeans trt/diff;
run;
```

Differences of Least Squares Means

Effect	trt	_trt	Estimate	Standard Error	DF	t Value	Pr > |t|
trt	1	2	0	0.7071	15	0.00	1.0000
trt	1	3	0	0.7071	15	0.00	1.0000
trt	1	4	0	0.7071	15	0.00	1.0000
trt	1	5	0	0.7071	15	0.00	1.0000
trt	1	6	0	0.7071	15	0.00	1.0000
trt	2	3	0	0.7071	15	0.00	1.0000
trt	2	4	0	0.7071	15	0.00	1.0000
trt	2	5	0	0.7071	15	0.00	1.0000
trt	2	6	0	0.7071	15	0.00	1.0000
trt	3	4	0	0.7071	15	0.00	1.0000
trt	3	5	0	0.7071	15	0.00	1.0000
trt	3	6	0	0.7071	15	0.00	1.0000
trt	4	5	0	0.7071	15	0.00	1.0000
trt	4	6	0	0.7071	15	0.00	1.0000
trt	5	6	0	0.7071	15	0.00	1.0000

The process of finding a balanced incomplete block design in this way is tedious. Also, this is certainly not the only trial-and-error method to find a good design. The main purpose of the Exercise was to exemplify the principles of blocking and balance. In practice, design software such as CycDesigN 4.0 can be used to find an optimal or near-optimal design.

Exercise 10

The blocks are as follows:

Block	1	2	3	4	5	6	7
Units 1	x	x	x	x	x	x	x
2	x	x	x	x	x	x	x
3	x	x	x	x	.	.	.
4	x	x	.	.	.	.	.

The x's represent units. In the following, x's will be moved within columns to align with the treatments allocated to them. This is done for better visibility of the treatment allocation.

Since standard errors involving comparisons with O need to be smaller, I guess that O needs to be present in each block. Also, blocks 1 and 2 obviously should be complete blocks. The remaining blocks will be incomplete and need to be filled up so that a maximum of balanced is preserved.

Units 1	O	O	O	O	O	O	O
2	A	A	A	.	A	.	.
3	B	B	.	B	.	B	.
4	C	C	C	C	.	.	C

```
data C5;
do trt=1 to 4;
  do block=1 to 7;
    input a$ @@;
    if a ne ' ' then do;
        y=1; output;
      end;
  end;
end;
datalines;
O     O     O     O     O     O     O
A     A     A     .     A     .     .
B     B     .     B     .     B     .
C     C     C     C     .     .     C
;
proc print data=c5; run;

proc mixed data=c5;
ods output diffs=diffs;
class trt block;
model y=trt block;
parms (1)/hold=1;
lsmeans trt/diff;
run;

data diffs;
set diffs;
vd=stderr**2;
run;
```

Effect	trt	_trt	Estimate	StdErr	DF	tValue	Probt	vd
trt	1	2	0	0.6660	10	0.00	1.0000	0.44361
trt	1	3	0	0.6660	10	0.00	1.0000	0.44361
trt	1	4	0	0.6094	10	0.00	1.0000	0.37143
trt	2	3	0	0.7947	10	0.00	1.0000	0.63158
trt	2	4	0	0.7275	10	0.00	1.0000	0.52932
trt	3	4	0	0.7275	10	0.00	1.0000	0.52932

The variances of differences (v.d.) are still a bit unbalanced, but the ratio of 2/3 of v.d. involving O and v.d. not involving O is is roughly achieved. Analysis with random blocks is as follows:

```
proc mixed data=c5;
ods output diffs=diffs2;
class trt block;
model y=trt;
random block;
parms (0.5)(1)/hold=1,2;
lsmeans trt/diff;
run;

data diffs2;
set diffs2;
vd2=stderr**2;
run;

proc print data=diffs2;
run;
```

Effect	trt	_trt	Estimate	StdErr	DF	tValue	Probt	vd2
trt	1	2	0	0.6469	10	0.00	1.0000	0.41845
trt	1	3	0	0.6469	10	0.00	1.0000	0.41845
trt	1	4	0	0.5975	10	0.00	1.0000	0.35706
trt	2	3	0	0.7506	10	0.00	1.0000	0.56338
trt	2	4	0	0.6987	10	0.00	1.0000	0.48818
trt	3	4	0	0.6987	10	0.00	1.0000	0.48818

Standard errors and variances of differences are somewhat reduced, because the inter-block information is now exploited.

Exercise 11

(1)

```
data c6;
input rep block trt yield;
yield=yield/100;
datalines;
1 1 1  505
1 1 2  587
1 1 3  470
1 1 4  519
1 2 5  472
1 2 6  415
1 2 7  580
1 2 8  478
1 3 9  489
1 3 10 525
1 3 11 446
1 3 12 513
1 4 13 468
1 4 14 486
1 4 15 476
1 4 16 473
2 1 1  478
2 1 5  453
2 1 9  449
2 1 13 512
2 2 2  496
2 2 6  329
2 2 10 400
2 2 14 432
2 3 3  474
2 3 7  573
2 3 11 424
2 3 15 450
2 4 4  516
2 4 8  497
2 4 12 457
2 4 16 467
3 1 1  468
3 1 6  431
3 1 11 461
3 1 16 437
3 2 2  417
3 2 7  455
3 2 12 513
3 2 13 494
```

```
3 3 3  417
3 3 8  455
3 3 9  466
3 3 14 434
3 4 4  478
3 4 5  461
3 4 10 458
3 4 15 376
;
proc glm data=c6;
class rep block trt;
model yield=rep rep*block trt;
run;
```

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	26	8.01276042	0.30818309	2.52	0.0172
Error	21	2.57256458	0.12250308		
Corrected Total	47	10.58532500			

R-Square	Coeff Var	Root MSE	yield Mean
0.756969	7.456818	0.350004	4.693750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
rep	2	1.54871250	0.77435625	6.32	0.0071
rep*block	9	1.91466250	0.21274028	1.74	0.1426
trt	15	4.54938542	0.30329236	2.48	0.0279

Source	DF	Type III SS	Mean Square	F Value	Pr > F
rep	2	1.54871250	0.77435625	6.32	0.0071
rep*block	9	1.44112292	0.16012477	1.31	0.2910
trt	15	4.54938542	0.30329236	2.48	0.0279

The Type III SS for rep and rep*block coincide with that given by Schuster and Lochow. However, the SS for trt is different. This is because GLM produces SS for trt that are adjusted for blocks and replicates, while Schuster and Lochow give unadjusted SS. This treatment SS is obtained by the Type I SS when fitting trt before rep and rep*block.

```
proc glm data=c6;
class rep block trt;
model yield=trt rep rep*block;
run;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	15	5.02292500	0.33486167	2.73	0.0172
rep	2	1.54871250	0.77435625	6.32	0.0071
rep*block	9	1.44112292	0.16012477	1.31	0.2910

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	15	4.54938542	0.30329236	2.48	0.0279
rep	2	1.54871250	0.77435625	6.32	0.0071
rep*block	9	1.44112292	0.16012477	1.31	0.2910

```
proc mixed data=c6;
class rep block trt;
model yield=rep rep*block trt;
run;
```

```
            The Mixed Procedure

       Type 3 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F

rep             2      21       6.32    0.0071
rep*block       9      21       1.31    0.2910
trt            15      21       2.48    0.0279
```

PROC MIXED produces the same F-values and p-values as GLM (Type III analysis), but it does not produce SS.

(2)

```
/*efficiency computation*/

/*get s.e.d. and v.d. for lattice analysis*/
proc mixed data=c6;
ods output diffs=diffs;
class rep block trt;
model yield=rep rep*block trt;
lsmeans trt/diff;
run;

data diffs;
set diffs;
vd_lattice=stderr**2;

proc means data=diffs;
var vd_lattice;
output out=vd_lattice mean=;
run;

/*get s.e.d. and v.d. for RCBD analysis*/
proc mixed data=c6;
ods output diffs=diffs2;
class rep trt;
model yield=rep trt;
lsmeans trt/diff;
run;

data diffs2;
set diffs2;
vd_RCBD=stderr**2;

proc means data=diffs2;
var vd_RCBD;
output out=vd_RCBD mean=;
run;

/*compute efficency*/
data efficiency;
merge vd_lattice vd_RCBD;
efficiency=vd_RCBD/vd_lattice;
```

```
run;

proc print data=efficiency;
run;
```

Obs	_TYPE_	_FREQ_	vd_lattice	vd_RCBD	efficiency
1	0	120	0.10617	0.089193	0.84010

The lattice design as an efficiency lower than one, so incomplete blocking was not worthwhile.

(3)

The design was taken from a basic plan without randomizing blocks within replicates and without randomizing treatments within blocks. This is not proper procedure.

(4)

No. The design is optimized so that the number of pairwise associations is as balanced as possible. Thus, the number of associations per pair is either zero or one.

Exercise 12

We need to fit a nested mixed model with random effects for samples nested within collection times and fished nested within sample. The effects are

colltime*sample + colltime*sample*fished
where the latter effect coincides with the residual error term and so does not need to be specified explicitly. Colltime is the fixed treatment effect (factor of interest). Thus, the code for MIXED is:

```
data c7;
input
colltime    sample     fishID  length;
datalines;
    1          1          1       5
    1          1          2       5
    1          1          3       5
    1          1          4       5
<more data>
    3          2        414       8
    3          2        415       9
    3          2        416       9
;
proc mixed data=c7;
class colltime sample;
model length=colltime / ddfm=KR;
random colltime*sample;
lsmeans colltime/pdiff;
run;
```

Covariance Parameter Estimates

Cov Parm	Estimate
colltime*sample	0
Residual	1.2484

The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	1278.4
AIC (smaller is better)	1280.4
AICC (smaller is better)	1280.4
BIC (smaller is better)	1280.2

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
colltime	2	413	9.93	<.0001

Least Squares Means

Effect	colltime	Estimate	Standard Error	DF	t Value	Pr > |t|
colltime	1	6.5652	0.1042	413	63.01	<.0001
colltime	2	6.2846	0.09799	413	64.13	<.0001
colltime	3	5.9708	0.08544	413	69.88	<.0001

Differences of Least Squares Means

Effect	colltime	_colltime	Estimate	Standard Error	DF	t Value	Pr > |t|
colltime	1	2	0.2806	0.1430	413	1.96	0.0505
colltime	1	3	0.5945	0.1347	413	4.41	<.0001
colltime	2	3	0.3139	0.1300	413	2.41	0.0162

The sample variance is estimated at zero. The F-test for colltime rejects the global null hypothesis that there are no differences among collection times in the average fish length. The comparison of collection times 1 and 2 is just not significant ($p = 0.0505$), while the other two comparisons are significant ($p < 0.0001$ and $p = 0.0162$). Inspection of the means indicates a linear decrease of length with time.

[The following matter is in square brackets because it was not specifically required in the exercise.]

[This can be verified by fitting a linear regression and testing for lack of fit (see Exercise C3). The code is as follows:

```
data c7;
set c7;
LOF=colltime;

proc mixed data=c7;
class LOF sample;
model length=colltime LOF/ ddfm=KR htype=1;
random colltime*sample;
run;
```

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
colltime	1	413	19.84	<.0001
LOF	1	413	0.02	0.8889

The lack of fit test is not significant (p = 0.8889), while the linear regression is significant (p < 0.0001). Note that the HTYPE=1 option was used to enforce sequential F-tests and that the LOF effect is fitted after the linear regression term. Also note that the regression variable COLLTIME was not listed in the CLASS statement, but the lack of fit variable LOF was lister in CLASS.]

Exercise 13

```
data c8;
input block$ main sub time$ species yield;
datalines;
d 1   1     I   1 31.2
d 1   2     I   3 10.5
d 1   3     I   2 15.6
d 2   4     III 3 13.8
d 2   5     III 2  9.8
d 2   6     III 1 10.5
d 3   7     II  2 24.2
d 3   8     II  1 26.8
d 3   9     II  3 14.8
c 4  10     III 2  8.5
c 4    11   III 3 15.3
c 4    12   III 1  9.8
c 5    13   I   3  8.5
c 5    14   I   1 28.9
c 5    15   I   2 13.0
c 6    16   II  1 24.3
c 6    17   II  2 25.5
c 6    18   II  3 14.2
b 7    19   II  3 18.2
b 7    20   II  1 29.2
b 7    21   II  2 28.1
b 8    22   III 1 12.0
b 8    23   III 2 13.0
b 8    24   III 3 18.5
b 9    25   I   2 17.3
b 9    26   I   3 12.3
b 9    27   I   1 35.2
a 10 28     I   1 30.5
a 10 29     I   2 15.1
a 10 30     I   3 10.0
a 11 31     II  2 27.1
a 11 32     II  3 26.9
```

```
a 11 33     II  1 15.5
a 12 34     III 3 10.3
a 12 35     III 1 12.5
a 12 36     III 2 17.6
;
proc mixed data=c8 nobound;
class block main sub time species;
model yield=block time species time*species /ddfm=satterthwaite;
random block*main;
run;

/*alternative solution*/
proc mixed data=c8 nobound;
class block main sub time species;
model yield=block time species time*species /ddfm=satterthwaite;
random block*time;
run;
```

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	3	6	20.73	0.0014
time	2	6	259.31	<.0001
species	2	18	12.49	0.0004
time*species	4	18	13.13	<.0001

There are significant interactions (F=13.13, p<0.0001), so we compare simple means. Note that the NOBOUND option was used to allow negative variance component estimates. In fact, in the present case the main plot error variance turns out to be negative. This is necessary in case of the split-plot design with balanced data in order to obtain exact tests, corresponding to the use of ANOVA estimators of variance components.

```
/*compare means*/
proc mixed data=c8 nobound;
ods output diffs=diffs lsmeans=lsmeans;
class block main sub time species;
model yield=block time species time*species /ddfm=satterthwaite;
random block*main;
lsmeans time*species/pdiff;
run;
```

The output is tedious and voluminous, because all comparisons are given, but we only need a portion of comparisons, where the level of one of the two factors is constant. ODS can be used to write differences into a SAS dataset and process this in a datastep.

Differences for constant level of time:

```
data diffs_by_time;
set diffs;
if time=_time;
run;

proc print data=diffs_by_time;
run;
```

Effect	time	species	_time	_species	Estimate	StdErr	DF	tValue	Probt
time*species	I	1	I	2	16.2000	2.6939	18	6.01	<.0001
time*species	I	1	I	3	21.1250	2.6939	18	7.84	<.0001
time*species	I	2	I	3	4.9250	2.6939	18	1.83	0.0841
time*species	II	1	II	2	-2.2750	2.6939	18	-0.84	0.4095
time*species	II	1	II	3	5.4250	2.6939	18	2.01	0.0592
time*species	II	2	II	3	7.7000	2.6939	18	2.86	0.0104
time*species	III	1	III	2	-1.0250	2.6939	18	-0.38	0.7080
time*species	III	1	III	3	-3.2750	2.6939	18	-1.22	0.2398
time*species	III	2	III	3	-2.2500	2.6939	18	-0.84	0.4145

Differences for constant level of species:

```
data diffs_by_species;
set diffs;
if species=_species;
run;

proc sort;
by species;

proc print data=diffs_by_species;
run;
```

Effect	time	species	_time	_species	Estimate	StdErr	DF	tValue	Probt
time*species	I	1	II	1	7.5000	2.2462	19.5	3.34	0.0034
time*species	I	1	III	1	20.2500	2.2462	19.5	9.02	<.0001
time*species	II	1	III	1	12.7500	2.2462	19.5	5.68	<.0001
time*species	I	2	II	2	-10.9750	2.2462	19.5	-4.89	<.0001
time*species	I	2	III	2	3.0250	2.2462	19.5	1.35	0.1935
time*species	II	2	III	2	14.0000	2.2462	19.5	6.23	<.0001
time*species	I	3	II	3	-8.2000	2.2462	19.5	-3.65	0.0016
time*species	I	3	III	3	-4.1500	2.2462	19.5	-1.85	0.0799
time*species	II	3	III	3	4.0500	2.2462	19.5	1.80	0.0869

There is a SAS macro to obtain a letter display that can be downloaded from our webpage at

https://www.uni-hohenheim.de/bioinformatik/beratung/toolsmacros/sasmacros/mult.sas

You need to copy the code and paste into a new program editor window in SAS. Then submit the whole window (or just the definition of the macro, (starting from %macro to %mend). Read the header of the macro for detailed instructions. Also, there are two worked examples at the end. **I will not require you to use the macro in an exam!**

Once this is done, the macro is available for use. To get comparisons for constant level of species, use this code.

```
%mult(trt=time, by=species, level=1);
%mult(trt=time, by=species, level=2);
%mult(trt=time, by=species, level=3);
```

trt	by	level	by2	level2	by3	level3	label	lsmean	g		
time	species	1	.	.	.	.	I	31.45	.	.	c
							II	23.95	.	b	.
							III	11.2	a	.	.

```
trt  by      level by2 level2 by3 level3 label    lsmean g

time species 2     .   .      .   .      I         15.25 a .
                                         II       26.225 . b
                                         III      12.225 a .

trt  by      level by2 level2 by3 level3 label    lsmean g

time species 3     .   .      .   .      I        10.325 a .
                                         II       18.525 . b
                                         III      14.475 a b
```

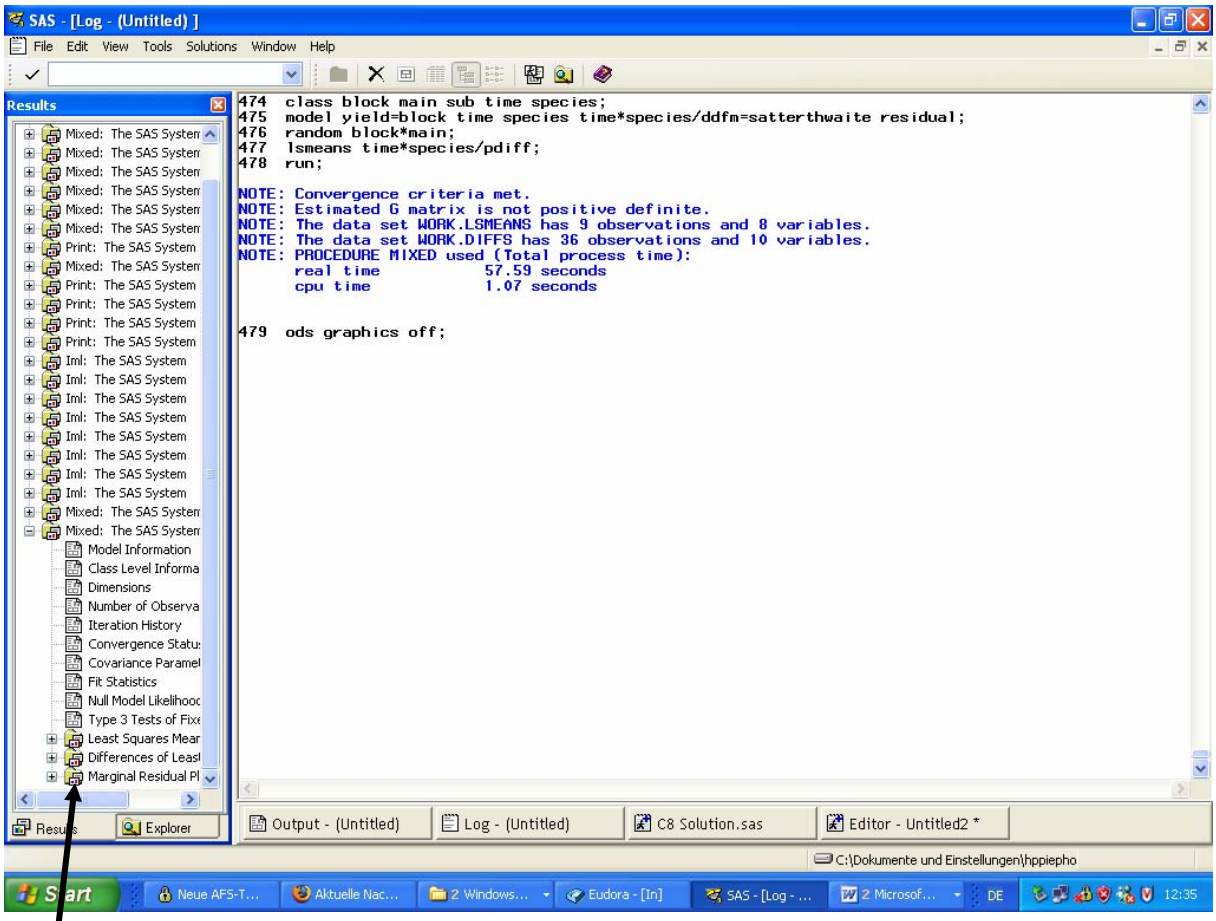
Means followed by a common letter are not significantly different at the 5% level according to a t-test. Comparisons for constant level of species:

```
%mult(trt=species, by=time, level='I');
%mult(trt=species, by=time, level='II');
%mult(trt=species, by=time, level='III');
```

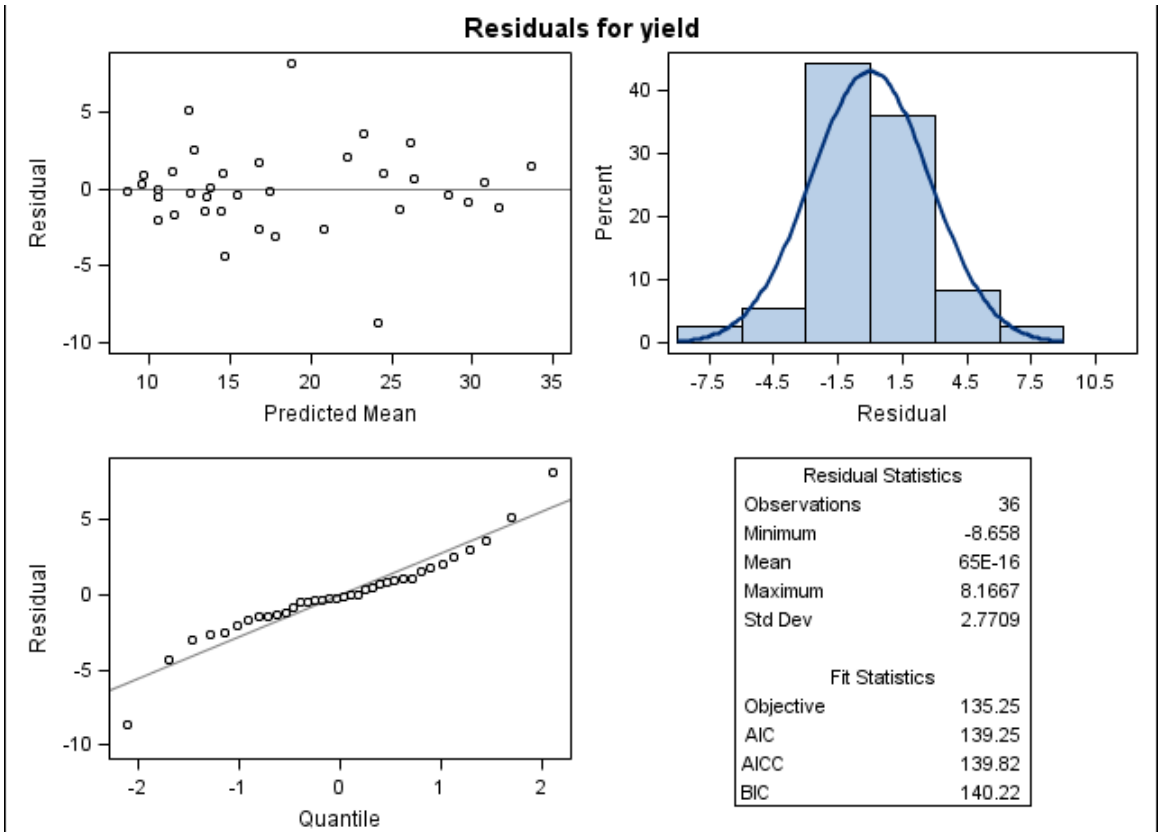
```
trt     by   level by2 level2 by3 level3     label    lsmean g

species time 'I'   .   .      .   .             1     31.45 . b
                                                 2     15.25 a .
                                                 3    10.325 a .

trt     by   level by2 level2 by3 level3     label    lsmean g

species time 'II'  .   .      .   .             1     23.95 a b
                                                 2    26.225 . b
                                                 3    18.525 a .


trt     by   level by2 level2 by3 level3     label    lsmean g

species time 'III' .   .      .   .             1      11.2 a
                                                 2    12.225 a
                                                 3    14.475 a
```

If you want to inspect residual plots, you can use the ODS GRAPHICS option as follows.

```
/*inspect residuals*/
ods graphics on;
proc mixed data=c8 nobound;
ods output diffs=diffs lsmeans=lsmeans;
class block main sub time species;
model yield=block time species time*species/ddfm=satterthwaite residual;
random block*main;
lsmeans time*species/pdiff;
run;
ods graphics off;
```



To get the residual plots, click on the appropriate icons in the results window. For example, the "marginal residuals" $\left(y - X\hat{\beta}\right)$ are as follows:



These residuals look inconspicuous.

Exercise 14

Plants are nested within pots. The model needs a random effect for pots to reflect the sub-sampling of plants within pots. Pots are coded within factors “temp” and “time”, so the pot effect must be coded by ”temp*time*pot” to fit a separate effect for each pot.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
temp	1	12	70.45	<.0001
time	2	12	5.18	0.0239
temp*time	2	12	1.32	0.3038

There is no interaction between “time” and “temp” (p=0.3038), so we compare marginal means for “temp” and “time”, which are significant with p<0.0001 and p=0.0239, respectively.

Least Squares Means

Effect	temp	time	Estimate	Standard Error	DF	t Value	Pr > |t|
temp	high		7.2361	0.2445	12	29.59	<.0001
temp	low		4.3333	0.2445	12	17.72	<.0001
time		8	5.5000	0.2995	12	18.36	<.0001
time		12	5.2917	0.2995	12	17.67	<.0001
time		16	6.5625	0.2995	12	21.91	<.0001

Differences of Least Squares Means

Effect	temp	time	_temp	_time	Estimate	Standard Error	DF	t Value	Pr > |t|
temp	**high**		**low**		**2.9028**	**0.3458**	**12**	**8.39**	**<.0001**
time		8		12	0.2083	0.4236	12	0.49	0.6317
time		**8**		**16**	**-1.0625**	**0.4236**	**12**	**-2.51**	**0.0275**
time		**12**		**16**	**-1.2708**	**0.4236**	**12**	**-3.00**	**0.0111**

Bold-faced comparisons are significant at α=0.05.

```
data a13;
input
temp$    time    pot    plant    stemgrowth;
datalines;
low       8       1       1           3.5
low       8       2       1           2.5
low       8       3       1           3.0
low      12       1       1           5.0
low      12       2       1           3.5
low      12       3       1           4.5
low      16       1       1           5.0
low      16       2       1           5.5
low      16       3       1           5.5
low       8       1       2           4.0
low       8       2       2           4.5
low       8       3       2           3.0
low      12       1       2           5.5
```

```
low      12      2       2          3.5
low      12      3       2          4.0
low      16      1       2          4.5
low      16      2       2          6.0
low      16      3       2          4.5
low       8      1       3          3.0
low       8      2       3          5.5
low       8      3       3          2.5
low      12      1       3          4.0
low      12      2       3          3.0
low      12      3       3          4.0
low      16      1       3          5.0
low      16      2       3          5.0
low      16      3       3          6.5
low       8      1       4          4.5
low       8      2       4          5.0
low       8      3       4          3.0
low      12      1       4          3.5
low      12      2       4          4.0
low      12      3       4          5.0
low      16      1       4          4.5
low      16      2       4          5.0
low      16      3       4          5.5
high      8      1       1          8.5
high      8      2       1          6.5
high      8      3       1          7.0
high     12      1       1          6.0
high     12      2       1          6.0
high     12      3       1          6.5
high     16      1       1          7.0
high     16      2       1          6.0
high     16      3       1         11.0
high      8      1       2          6.0
high      8      2       2          7.0
high      8      3       2          7.0
high     12      1       2          5.5
high     12      2       2          8.5
high     12      3       2          6.5
high     16      1       2          9.0
high     16      2       2          7.0
high     16      3       2          7.0
high      8      1       3          9.0
high      8      2       3          8.0
high      8      3       3          7.0
high     12      1       3          3.5
high     12      2       3          4.5
high     12      3       3          8.5
high     16      1       3          8.5
high     16      2       3          7.0
high     16      3       3          9.0
high      8      1       4          8.5
high      8      2       4          6.5
high      8      3       4          7.0
high     12      1       4          7.0
high     12      2       4          7.5
high     12      3       4          7.5
high     16      1       4          8.5
high     16      2       4          7.0
high     16      3       4          8.0
;
proc mixed data=a13;
class temp time pot plant;
model stemgrowth=temp time temp*time/ddfm=KR;
```

```
lsmeans temp time/pdiff; /*interaction n.s.*/
random temp*time*pot;
run;
```

Exercise 15

Date is the main plot factor of this split-plot experiment. In order to do the linear regression on time, we need to define a time variable. Without loss of generality, we assess the time variable in day units and define Oct. 15 as day=0. Hence, Nov. 1 is day=17 and Nov.15 is day=31.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	2	4	0.22	0.8113
day	0	.	.	.
V	3	18	23.89	<.0001
day*V	0	.	.	.
V*lackfit	4	13.6	11.40	0.0003

There is a significant lack of fit for linear time trend (p=0.0003).

```
data a12;
input
block D V yield;
datalines;
1 2 2 24
1 2 3 19
1 2 1 30
1 2 4 15
1 1 1 25
1 1 4 11
1 1 2 19
1 1 3 22
1 3 1 17
1 3 3 12
1 3 4  8
1 3 2 20
2 1 4 14
2 1 1 31
2 1 3 20
2 1 2 14
2 2 2 20
2 2 3 18
2 2 1 32
2 2 4 13
2 3 4 13
2 3 1 20
2 3 2 16
2 3 3 17
3 1 1 28
3 1 3 17
3 1 2 16
3 1 4 14
3 3 4  8
3 3 3 15
3 3 1 19
3 3 2 20
```

```
3 2 3 16
3 2 1 28
3 2 2 24
3 2 4 19
;
data b12;
set a12;
if D=1 then day=0;
if D=2 then day=17;
if D=3 then day=31;
lackfit=D;

proc mixed data=b12 nobound;
class block D V lackfit;
model yield=block day V day*V lackfit*V/ddfm=kr;
random block*D;
run;
```

Exercise 16

It is a split plot design with main plot factor A randomized in complete blocks and sub plot factor B completely randomized within main plots.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block	3	3	5.16	0.1055
A	1	3	38.32	0.0085
B	2	12	102.23	<.0001
A*B	2	12	88.13	<.0001

The interaction is significant (p<0.001), so we compare a*b means.

Least Squares Means

Effect	A	B	Estimate	Standard Error
A*B	1	1	22.9500	0.6137
A*B	1	2	19.1750	0.6137
A*B	1	3	28.6500	0.6137
A*B	2	1	27.9500	0.6137
A*B	2	2	28.4250	0.6137
A*B	2	3	28.7000	0.6137

Differences of Least Squares Means

Effect	A	B	_A	_B	Estimate	Standard Error	DF	t Value	Pr > |t|
A*B	**1**	**1**	**1**	**2**	**3.7750**	**0.4905**	**12**	**7.70**	**<.0001**
A*B	**1**	**1**	**1**	**3**	**-5.7000**	**0.4905**	**12**	**-11.62**	**<.0001**
A*B	**1**	**1**	**2**	**1**	**-5.0000**	**0.8679**	**4.76**	**-5.76**	**0.0026**
A*B	**1**	**1**	**2**	**2**	**-5.4750**	**0.8679**	**4.76**	**-6.31**	**0.0018**
A*B	**1**	**1**	**2**	**3**	**-5.7500**	**0.8679**	**4.76**	**-6.62**	**0.0014**
A*B	**1**	**2**	**1**	**3**	**-9.4750**	**0.4905**	**12**	**-19.32**	**<.0001**
A*B	**1**	**2**	**2**	**1**	**-8.7750**	**0.8679**	**4.76**	**-10.11**	**0.0002**
A*B	**1**	**2**	**2**	**2**	**-9.2500**	**0.8679**	**4.76**	**-10.66**	**0.0002**
A*B	**1**	**2**	**2**	**3**	**-9.5250**	**0.8679**	**4.76**	**-10.97**	**0.0001**

A*B	1	3	2	1	0.7000	0.8679	4.76	0.81	0.4584
A*B	1	3	2	2	0.2250	0.8679	4.76	0.26	0.8063
A*B	1	3	2	3	-0.05000	0.8679	4.76	-0.06	0.9564
A*B	2	1	2	2	-0.4750	0.4905	12	-0.97	0.3519
A*B	2	1	2	3	-0.7500	0.4905	12	-1.53	0.1521
A*B	2	2	2	3	-0.2750	0.4905	12	-0.56	0.5853

Boldfaced comparisons are significant. But not all of these are relevant. We may drop all comparisons, where both A and B are varied. This leaves the following comparisons:

Conditioning on A:

Differences of Least Squares Means

Effect	A	B	_A	_B	Estimate	Standard Error	DF	t Value	Pr > |t|
A*B	**1**	**1**	**1**	**2**	**3.7750**	**0.4905**	**12**	**7.70**	**<.0001**
A*B	**1**	**1**	**1**	**3**	**-5.7000**	**0.4905**	**12**	**-11.62**	**<.0001**
A*B	**1**	**2**	**1**	**3**	**-9.4750**	**0.4905**	**12**	**-19.32**	**<.0001**
A*B	2	1	2	2	-0.4750	0.4905	12	-0.97	0.3519
A*B	2	1	2	3	-0.7500	0.4905	12	-1.53	0.1521
A*B	2	2	2	3	-0.2750	0.4905	12	-0.56	0.5853

Levels of B only differ significantly for A=1.

Conditioning on B:

Differences of Least Squares Means

Effect	A	B	_A	_B	Estimate	Standard Error	DF	t Value	Pr > |t|
A*B	**1**	**1**	**2**	**1**	**-5.0000**	**0.8679**	**4.76**	**-5.76**	**0.0026**
A*B	**1**	**2**	**2**	**2**	**-9.2500**	**0.8679**	**4.76**	**-10.66**	**0.0002**
A*B	1	3	2	3	-0.05000	0.8679	4.76	-0.06	0.9564

Levels of A differ significantly only for B=1 and B=2.

```
data a10;
input
block A B yield;
datalines;
1 2 2 25.5
1 2 1 24.9
1 2 3 25.8
1 1 3 26.1
1 1 2 18.0
1 1 1 21.7
2 1 1 21.1
2 1 2 17.9
2 1 3 28.9
2 2 1 27.6
2 2 3 29.4
2 2 2 29.3
3 1 3 28.6
3 1 2 19.5
3 1 1 23.2
3 2 2 29.7
3 2 3 30.3
3 2 1 29.5
```

```
4 2 3 29.3
4 2 2 29.2
4 2 1 29.8
4 1 2 21.3
4 1 3 31.0
4 1 1 25.8
;
proc mixed data=a10;
class block a b;
model yield=block a b a*b/ddfm=KR;
random block*a;
lsmeans a*b/pdiff; /*interaction significant!*/
run;
```

Exercise 17

First select covariance structure with saturated model.

Model	AIC
UNR	194.3
AR(1)	200.5
CS	198.5
ARH(1)	192.0
CSH	188.1

The CSH model fits best, so we use this to test fixed effects.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block*year	12	14.4	55.93	<.0001
P	1	10.1	23.96	0.0006
K	1	10.1	0.24	0.6332
P*K	1	10.1	0.13	0.7227
time*P	0	.	.	.
time*K	0	.	.	.
time*P*K	0	.	.	.
P*K*lackfit	6	17.3	2.95	0.0364

There is a significant lack of fit for a linear trend model, so try the quadratic.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block*year	12	14.4	55.93	<.0001
P	1	10.4	21.31	0.0009
K	1	10.4	0.85	0.3787
P*K	1	10.4	0.33	0.5805
time*P	0	.	.	.
time*K	0	.	.	.
time*P*K	0	.	.	.
time*time*P	0	.	.	.
time*time*K	0	.	.	.
time*time*P*K	0	.	.	.
P*K*lackfit	3	22.3	1.01	0.4088

The quadratic model fits fine, so we use this to test for heterogeneity.

Type 1 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
block*year	15	16.4	52.35	<.0001
P	1	13.1	183.20	<.0001
K	1	13.1	13.13	0.0031
P*K	1	13.1	0.04	0.8508
time	0	.	.	.
time*P	1	20	7.52	0.0126
time*K	1	20	1.40	0.2506
time*P*K	1	20	0.20	0.6594
time*time*P	**1**	**20.8**	**12.96**	**0.0017**
time*time*K	1	20.8	0.53	0.4732
time*time*P*K	1	20.8	0.38	0.5431

There is significant heterogeneity in the curves only between levels of P (p=0.0017).

```
data a11;
input
  block    year    K    P    yield;
datalines;
     1         1     0    0    24.25
     1         1     0    1    43.50
     1         1     1    0    26.00
     1         1     1    1    37.75
     1         2     0    0    43.50
     1         2     0    1    47.75
     1         2     1    0    43.50
     1         2     1    1    48.25
     1         3     0    0    38.50
     1         3     0    1    43.50
     1         3     1    0    42.75
     1         3     1    1    46.50
     1         4     0    0    41.50
     1         4     0    1    42.25
     1         4     1    0    45.50
     1         4     1    1    47.35
     2         1     0    0    38.75
     2         1     0    1    42.25
     2         1     1    0    43.50
     2         1     1    1    48.50
     2         2     0    0    47.00
     2         2     0    1    48.50
     2         2     1    0    46.50
     2         2     1    1    50.75
     2         3     0    0    54.25
     2         3     0    1    56.75
     2         3     1    0    56.50
     2         3     1    1    59.50
     2         4     0    0    46.00
     2         4     0    1    51.75
     2         4     1    0    45.75
     2         4     1    1    50.25
     3         1     0    0    22.75
     3         1     0    1    34.00
     3         1     1    0    22.25
     3         1     1    1    37.00
     3         2     0    0    41.75
```

```
     3         2     0     1    48.00
     3         2     1     0    42.00
     3         2     1     1    45.00
     3         3     0     0    42.25
     3         3     0     1    39.75
     3         3     1     0    42.75
     3         3     1     1    42.25
     3         4     0     0    40.75
     3         4     0     1    46.25
     3         4     1     0    43.00
     3         4     1     1    49.75
     4         1     0     0    33.50
     4         1     0     1    46.50
     4         1     1     0    40.50
     4         1     1     1    49.00
     4         2     0     0    45.75
     4         2     0     1    48.00
     4         2     1     0    45.00
     4         2     1     1    48.50
     4         3     0     0    53.25
     4         3     0     1    54.75
     4         3     1     0    53.25
     4         3     1     1    57.75
     4         4     0     0    42.00
     4         4     0     1    49.25
     4         4     1     0    46.50
     4         4     1     1    52.25
;
/*UNR*/
proc mixed data=a11 lognote;
class P K block year;
model yield=block*year P*year K*year P*K*year/ddfm=kr;
repeated year / subject=block*P*K type=unr;
run;

/*AR(1)*/
proc mixed data=a11;
class P K block year;
model yield=block*year P*year K*year P*K*year/ddfm=kr;
repeated year / subject=block*P*K type=ar(1);
run;

/*CS*/
proc mixed data=a11;
class P K block year;
model yield=block*year P*year K*year P*K*year/ddfm=kr;
repeated year / subject=block*P*K type=CS;
run;

/*ARH(1)*/
proc mixed data=a11;
class P K block year;
model yield=block*year P*year K*year P*K*year/ddfm=kr;
repeated year / subject=block*P*K type=arh(1);
run;

/*CSH*/
proc mixed data=a11;
class P K block year;
model yield=block*year P*year K*year P*K*year/ddfm=kr;
repeated year / subject=block*P*K type=CSH;
run;
```

```
/*lack of fit for linear model, using CSH*/
/*CSH*/

data b11;
set a11;
time=year;
lackfit=year;
run;

proc mixed data=b11;
class P K block year lackfit;
model yield=block*year P K P*K
                       P*time K*time P*K*time P*K*lackfit/ddfm=kr;
repeated year / subject=block*P*K type=CSH;
run;

/*linear model shows lack of fit, so we try quadratic*/

proc mixed data=b11;
class P K block year lackfit;
model yield=block*year P K P*K
                       P*time K*time P*K*time
                       P*time*time K*time*time P*K*time*time
                       P*K*lackfit/ddfm=kr;
repeated year / subject=block*P*K type=CSH;
run;

/*quadratic model fits, so we use this to test heterogeneity*/

proc mixed data=b11;
class P K block year lackfit;
model yield=block*year P K P*K time
                       P*time K*time P*K*time
                       P*time*time K*time*time P*K*time*time
                       /ddfm=kr htype=1;
repeated year / subject=block*P*K type=CSH;
run;
```

Exercise 18

Model (Tree+Apple)	AIC
AR(1)	-1918.9
CS	-1837.4
ARH(1)	-1932.1
CSH	-1889.7
UN/UNR	did not converge

ARH(1) is preferred because it has the smallest AIC. Inspecting the parameter estimates, it is found that tree effects have very small variance. So we drop these effects from the model to see if fit is improved.

Model (Tree only)	AIC
ARH(1)	-1935.3
UN/UNR	-1959.2

This reduction, in fact, improved the fit. Moreover, we can now also get the UNR model to converge! And UNR fits best among all models!

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
t	0	.	.	.
lackfit	4	69.2	24.53	<.0001

The lack of fit for the linear model is significant, so we add the quadratic term.

Type 3 Tests of Fixed Effects

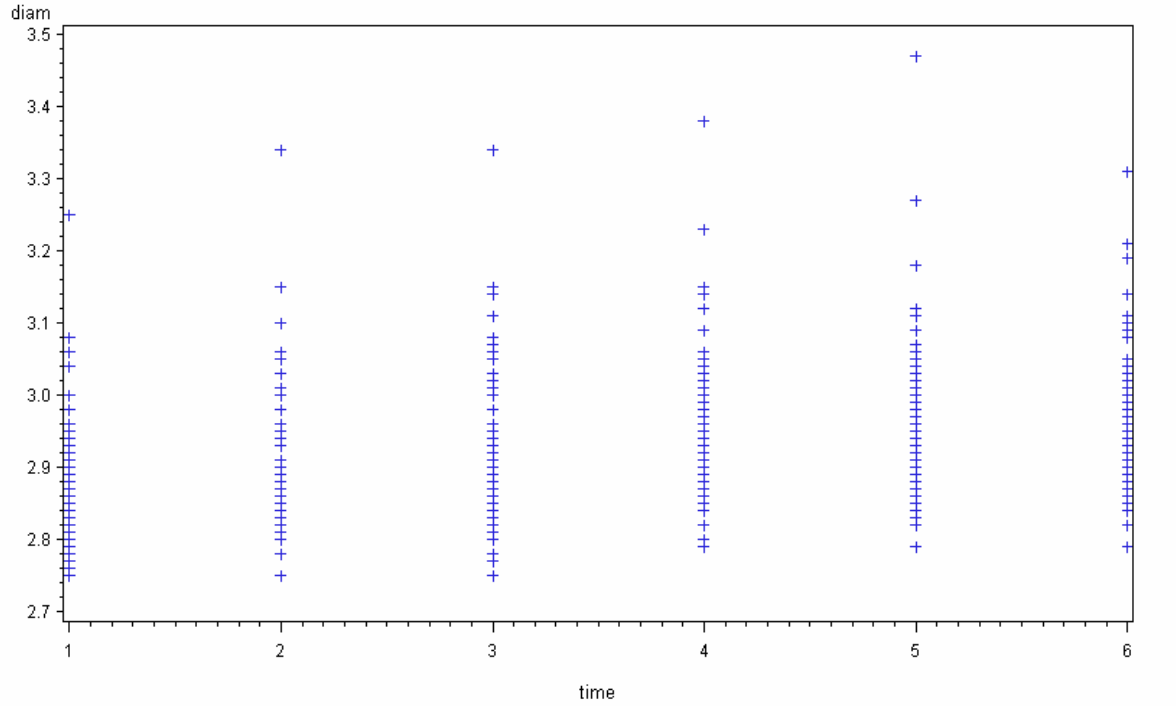
Effect	Num DF	Den DF	F Value	Pr > F
t	0	.	.	.
t*t	0	.	.	.
lackfit	3	71.5	5.86	0.0012

The lack of fit is still significant, so we try the cubic. This does not run with UNR, so we switch to ARH(1).

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
t	0	.	.	.
t*t	0	.	.	.
t*t*t	0	.	.	.
lackfit	2	250	8.69	0.0002

This does not work either, so we must conclude that a different model is needed. Inspecting the plot of diam versus time, it seems that diam increases to some asymptotic value, so a polynomial dies not seem a good idea. We stop here, but point out that some nonlinear regression model could be tried.




```
PROC IMPORT OUT= WORK.A15
            DATAFILE= "D:\e\hpp\Module\Bioinformatics\data\AppleData.xls"
            DBMS=EXCEL REPLACE;
     RANGE="SIZE7";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

proc gplot data=a15;
plot diam*time;
run;

/*AR(1)*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
random time/sub=tree type=AR(1);
repeated time/sub=tree*apple type=ar(1);
run;

/*CS*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
random time/sub=tree type=CS;
repeated time/sub=tree*apple type=CS;
run;

/*ARH(1)*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
random time/sub=tree type=ARH(1);
repeated time/sub=tree*apple type=arH(1);
run;
```

```
/*CSH*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
random time/sub=tree type=CSH;
repeated time/sub=tree*apple type=CSH;
run;

/*UN*/
proc mixed data=a15 lognote maxiter=1000;
class time tree apple;
model diam=time;
random time/sub=tree type=UN;
repeated time/sub=tree*apple type=UN;
run;

/*ARH(1) for apples only*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
repeated time/sub=tree*apple type=arH(1);
run;

/*UNR for apples only*/
proc mixed data=a15 lognote;
class time tree apple;
model diam=time;
repeated time/sub=tree*apple type=UNR;
run;

/*test trends using UNR for apples only*/
data b15;
set a15;
t=time;
lackfit=time;

/*lack of fit for linear model*/
proc mixed data=b15 lognote;
class time tree apple lackfit;
model diam=t lackfit/ddfm=kr;
repeated time/sub=tree*apple type=UNR;
run;

/*lack of fit for quadratic model*/
proc mixed data=b15 lognote;
class time tree apple lackfit;
model diam=t t*t lackfit/ddfm=kr;
repeated time/sub=tree*apple type=unr;
run;

/*lack of fit for cubic model*/
proc mixed data=b15 lognote;
class time tree apple lackfit;
model diam=t t*t t*t*t lackfit/ddfm=kr;
repeated time/sub=tree*apple type=arh(1);
run;
```

Exercise 19

Model	AIC
AR(1)	484.0
CS	483.6
ARH(1)	454.1
CSH	442.0
UNR	441.8

The unstructured model fits best. So we test the lack of fit for linear trend using this variance-covariance model.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
time	0	.	.	.
grp	4	25	9.84	<.0001
t*grp	0	.	.	.
grp*lackfit	4	25	2.74	0.0512

There is no significant lack of fit (p=0.0512).

```
data lemna;
input
grp    ind    time       y       count;
datalines;
     1       11       1      14.1497       26
     1       11       5      17.2428       53
     1       11       7      18.5126       71
     1       12       1      14.6240       29
     1       12       5      17.7085       59
     1       12       7      19.6379       92
     1       13       1      13.8021       24
     1       13       5      17.4036       55
     1       13       7      19.4939       89
     1       14       1      13.8021       24
     1       14       5      17.1600       52
     1       14       7      19.1908       83
     1       15       1      14.4716       28
     1       15       5      17.5587       57
     1       15       7      19.6848       93
     1       16       1      13.0103       20
     1       16       5      16.0206       40
     1       16       7      17.9239       62
     2       21       1      13.9794       25
     2       21       5      16.8124       48
     2       21       7      18.1954       66
     2       22       1      14.4716       28
     2       22       5      17.0757       51
     2       22       7      18.3885       69
     2       23       1      14.3136       27
     2       23       5      17.0757       51
     2       23       7      18.3251       68
     2       24       1      13.9794       25
     2       24       5      16.4345       44
     2       24       7      18.0618       64
```

```
     2       25       1      13.2222       21
     2       25       5      15.9106       39
     2       25       7      17.6343       58
     2       26       1      14.3136       27
     2       26       5      16.7210       47
     2       26       7      18.0618       64
     3       31       1      13.6173       23
     3       31       5      16.4345       44
     3       31       7      18.4510       70
     3       32       1      13.4242       22
     3       32       5      16.2325       42
     3       32       7      17.7815       60
     3       33       1      13.9794       25
     3       33       5      16.8124       48
     3       33       7      18.5126       71
     3       34       1      13.4242       22
     3       34       5      16.0206       40
     3       34       7      18.0618       64
     3       35       1      13.2222       21
     3       35       5      16.0206       40
     3       35       7      17.6343       58
     3       36       1      13.2222       21
     3       36       5      16.4345       44
     3       36       7      17.8533       61
     4       41       1      12.7875       19
     4       41       5      16.5321       45
     4       41       7      18.1954       66
     4       42       1      13.0103       20
     4       42       5      17.5587       57
     4       42       7      18.6923       74
     4       43       1      14.1497       26
     4       43       5      16.7210       47
     4       43       7      19.0309       80
     4       44       1      13.4242       22
     4       44       5      16.7210       47
     4       44       7      18.6923       74
     4       45       1      12.7875       19
     4       45       5      16.8124       48
     4       45       7      17.9934       63
     4       46       1      13.6173       23
     4       46       5      16.8124       48
     4       46       7      18.5126       71
     5       51       1      13.6173       23
     5       51       5      16.9897       50
     5       51       7      19.4448       88
     5       52       1      12.7875       19
     5       52       5      16.6276       46
     5       52       7      18.7506       75
     5       53       1      14.7712       30
     5       53       5      17.5587       57
     5       53       7      20.2119      105
     5       54       1      13.2222       21
     5       54       5      17.3239       54
     5       54       7      19.0849       81
     5       55       1      13.6173       23
     5       55       5      17.4819       56
     5       55       7      19.4448       88
     5       56       1      13.9794       25
     5       56       5      17.7085       59
     5       56       7      19.3952       87
;
proc gplot;
plot count*time=grp;
```

```
run;

/*AR(1)*/
proc mixed data=lemna;
class time grp ind;
model count=time grp time*grp;
repeated time / subject=ind type=AR(1);
run;

/*CS*/
proc mixed data=lemna;
class time grp ind;
model count=time grp time*grp;
repeated time / subject=ind type=CS;
run;

/*ARH(1)*/
proc mixed data=lemna;
class time grp ind;
model count=time grp time*grp;
repeated time / subject=ind type=ARH(1);
run;

/*CSH*/
proc mixed data=lemna;
class time grp ind;
model count=time grp time*grp;
repeated time / subject=ind type=CSH;
run;

/*UNR*/
proc mixed data=lemna;
class time grp ind;
model count=time grp time*grp;
repeated time / subject=ind type=UNR;
run;

/*test lack of fit*/
data lemna2;set lemna;
t=time;
lackfit=time;

proc mixed data=lemna2;
class time grp ind lackfit;
model count=grp t*grp lackfit*grp/ddfm=KR;
repeated time / subject=ind type=UNR;
run;
```

Exercise 20

Fitting models to this dataset is time consuming!

Different variance-covariance models for repeated measures cause problems of convergence. One way to tackle these is to modify starting values using the PARMS statement. The default used by MIXED is to initialize the residual at 1 and all other parameters at 0. By trial and error we may try to find better starting values. This is a thorny route, and no simple rules can be given. My own experience is that initializing all parameters at 1 often helps. But parameter restrictions must be observed. For example, correlations must be bounded above at 1 (UNR, AR(1)). The use of the PARMS statement is exemplified in the SAS code given below, where I give the final set of parameter values that worked for me.

The UNR did not converge properly. I fitted a so-called **factor-analytic model** with two factors instead, which in this case is a one-to-one re-parameterization of the UNR model that is guaranteed to yield a feasible (at least positive semi-definite) estimate of the variance-covariance matrix. It is defined as

$$\Sigma = \Lambda\Lambda^T + I\sigma^2,$$

where

$\Lambda = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{pmatrix}$ and Λ^T denotes the transpose of Λ. Note that the model has six parameters, just as the unstructured model.

Model	AIC
CS	-2964.5
AR(1)	-3084.2
UNR	did not converge
FA(2)	-3155.7

We use the FA(2) model to test linear trend for lack of fit.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
rep*time	6	99	1.48	0.1936
t*genotype	0	.	.	.
genotype*lackfit	566	1543	2.48	<.0001

There is a significant lack of fit (p<0.0001).

```
PROC IMPORT OUT= WORK.a7
            DATAFILE= "D:\e\hpp\Module\Bioinformatics\
                       data\arabidopsis.xls"
            DBMS=EXCEL REPLACE;
     RANGE="Tabelle1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

data b7;
set a7;
time=15; y=RD15; output;
time=22; y=RD22; output;
time=29; y=RD29; output; run;

data b7;
set b7;
logy=log(y); run;
```

```
/*compound symmetry*/
proc mixed data=b7 lognote;
class genotype rep block mainplot subplot time;
model logy=genotype*time rep*time;
random time/sub=rep*block type=cs;
random time/sub=rep*block*mainplot type=cs;
repeated time/sub=rep*block*mainplot*subplot type=cs;
run;


/*AR(1)*/
proc mixed data=b7 lognote;
class genotype rep block mainplot subplot time;
model logy=genotype*time rep*time;
random time/sub=rep*block type=ar(1);
random time/sub=rep*block*mainplot type=ar(1);
repeated time/sub=rep*block*mainplot*subplot type=ar(1);
parms (1)(.1)(1)(.1)(.1)(1);
run;

/*UNR - did not converge*/
proc mixed data=b7 lognote;
class genotype rep block mainplot subplot time;
model logy=genotype*time rep*time;
random time/sub=rep*block type=UNR;
random time/sub=rep*block*mainplot type=UNR;
repeated time/sub=rep*block*mainplot*subplot type=UNR;
parms (1)(1)(1)(.1)(.1)(.1)
      (1)(1)(1)(.1)(.1)(.1)
      (1)(1)(1)(.1)(.1)(.1)
;
run;

/*FA(2) - a reparameterization of UNR*/
ods output covparms=p; /*save final estimates for
                         use in next call of FA(2)*/
proc mixed data=b7 lognote;
class rep block mainplot subplot genotype time;
model logy=rep*time genotype*time;
random   time/sub=rep*block type=fa(2);
random   time/sub=rep*block*mainplot type=fa(2);
repeated time/sub=rep*block*mainplot*subplot type=fa(2);
run;

/*test lack of fit for linear model using FA(2)*/
data c7;
set b7;
lackfit=time;
t=time;

proc mixed data=c7 lognote;
class rep block mainplot subplot genotype time lackfit;
model logy=rep*time genotype*t genotype*lackfit;
random   time/sub=rep*block type=fa(2);
random   time/sub=rep*block*mainplot type=fa(2);
repeated time/sub=rep*block*mainplot*subplot type=fa(2);
parms / pdata=p; /*use estimates from last run as starting values*/
run;
```

Exercise 21

$$\boldsymbol{D}=\begin{pmatrix}\sigma_1^2 & 0 & 0\\ 0 & \sigma_2^2 & 0\\ 0 & 0 & \sigma_3^2\end{pmatrix},\ \boldsymbol{D}^{1/2}=\begin{pmatrix}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\\ 0 & 0 & \sigma_3\end{pmatrix},\ \boldsymbol{C}=\begin{pmatrix}1 & \rho & \rho^2\\ \rho & 1 & \rho\\ \rho^2 & \rho & 1\end{pmatrix},$$

$$\boldsymbol{D}^{1/2}\boldsymbol{C}=\begin{pmatrix}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\\ 0 & 0 & \sigma_3\end{pmatrix}\begin{pmatrix}1 & \rho & \rho^2\\ \rho & 1 & \rho\\ \rho^2 & \rho & 1\end{pmatrix}=\begin{pmatrix}\sigma_1 & \rho\sigma_1 & \rho^2\sigma_1\\ \rho\sigma_2 & \sigma_2 & \rho\sigma_2\\ \rho^2\sigma_3 & \rho\sigma_3 & \sigma_3\end{pmatrix}$$

$$\boldsymbol{D}^{1/2}\boldsymbol{C}\boldsymbol{D}^{1/2}=\begin{pmatrix}\sigma_1 & \rho\sigma_1 & \rho^2\sigma_1\\ \rho\sigma_2 & \sigma_2 & \rho\sigma_2\\ \rho^2\sigma_3 & \rho\sigma_3 & \sigma_3\end{pmatrix}\begin{pmatrix}\sigma_1 & 0 & 0\\ 0 & \sigma_2 & 0\\ 0 & 0 & \sigma_3\end{pmatrix}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3\\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3\\ \rho^2\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2\end{pmatrix}$$

Exercise 22

(1)

Reading the data:

```
data C9;
input
week    trt    rat    weight;
datalines;
      1       1       1       57
      2       1       1       86
      3       1       1      114
      4       1       1      139
      5       1       1      172
      1       1       2       60
      2       1       2       93
      3       1       2      123
      4       1       2      146
      5       1       2      177
      1       1       3       52
      2       1       3       77
      3       1       3      111
      4       1       3      144
      5       1       3      185
      1       1       4       49
      2       1       4       67
      3       1       4      100
      4       1       4      129
      5       1       4      164
      1       1       5       56
      2       1       5       81
      3       1       5      104
      4       1       5      121
      5       1       5      151
      1       1       6       46
      2       1       6       70
      3       1       6      102
      4       1       6      131
```

5	1	6	153
1	1	7	51
2	1	7	71
3	1	7	94
4	1	7	110
5	1	7	141
1	1	8	63
2	1	8	91
3	1	8	112
4	1	8	130
5	1	8	154
1	1	9	49
2	1	9	67
3	1	9	90
4	1	9	112
5	1	9	140
1	1	10	57
2	1	10	82
3	1	10	110
4	1	10	139
5	1	10	169
1	2	1	59
2	2	1	85
3	2	1	121
4	2	1	156
5	2	1	191
1	2	2	54
2	2	2	71
3	2	2	90
4	2	2	110
5	2	2	138
1	2	3	56
2	2	3	75
3	2	3	108
4	2	3	151
5	2	3	189
1	2	4	59
2	2	4	85
3	2	4	116
4	2	4	148
5	2	4	177
1	2	5	57
2	2	5	72
3	2	5	97
4	2	5	120
5	2	5	144
1	2	6	52
2	2	6	73
3	2	6	97
4	2	6	116
5	2	6	140
1	2	7	52
2	2	7	70
3	2	7	105
4	2	7	138
5	2	7	171
1	3	1	61
2	3	1	86
3	3	1	109
4	3	1	120
5	3	1	129
1	3	2	59
2	3	2	80

```
      3       3       2       101
      4       3       2       111
      5       3       2       126
      1       3       3        53
      2       3       3        79
      3       3       3       100
      4       3       3       106
      5       3       3       133
      1       3       4        59
      2       3       4        88
      3       3       4       100
      4       3       4       111
      5       3       4       122
      1       3       5        51
      2       3       5        75
      3       3       5       101
      4       3       5       123
      5       3       5       140
      1       3       6        51
      2       3       6        75
      3       3       6        92
      4       3       6       100
      5       3       6       119
      1       3       7        56
      2       3       7        78
      3       3       7        95
      4       3       7       103
      5       3       7       108
      1       3       8        58
      2       3       8        69
      3       3       8        93
      4       3       8       114
      5       3       8       138
      1       3       9        46
      2       3       9        61
      3       3       9        78
      4       3       9        90
      5       3       9       107
      1       3      10        53
      2       3      10        72
      3       3      10        89
      4       3      10       104
      5       3      10       122
;
```

The unstructured model is fitted by:

```
/*unstructured model*/
proc mixed data=c9;
class week rat trt;
model weight=week trt week*trt;
repeated week/subject=rat*trt type=un;
run;
```

```
                    Covariance Parameter Estimates

Cov Parm    Subject    Estimate

UN(1,1)     rat*trt     21.5756
UN(2,1)     rat*trt     33.0196
UN(2,2)     rat*trt     68.7274
UN(3,1)     rat*trt     31.5821
```

```
UN(3,2)     rat*trt     69.0607
UN(3,3)     rat*trt     94.7690
UN(4,1)     rat*trt     29.3762
UN(4,2)     rat*trt     64.5435
UN(4,3)     rat*trt      116.36
UN(4,4)     rat*trt      181.56
UN(5,1)     rat*trt     25.3774
UN(5,2)     rat*trt     57.3369
UN(5,3)     rat*trt      123.75
UN(5,4)     rat*trt      207.68
UN(5,5)     rat*trt      268.34
```

Fit Statistics

-2 Res Log Likelihood	737.3
AIC (smaller is better)	**767.3**
AICC (smaller is better)	771.9
BIC (smaller is better)	786.7

The compound symmetry (CS) model:

```
/*compound symmetry model*/
proc mixed data=c9;
class week rat trt;
model weight=week trt week*trt;
repeated week/subject=rat*trt type=cs;
run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	rat*trt	75.8086
Residual		51.1857

The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	896.6
AIC (smaller is better)	**900.6**
AICC (smaller is better)	900.8
BIC (smaller is better)	903.2

The AR(1) model:

```
/*AR(1) model*/
proc mixed data=c9;
class week rat trt;
model weight=week trt week*trt;
repeated week/subject=rat*trt type=AR(1);
run;
```

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
AR(1)	rat*trt	0.8830
Residual		137.55

```
            Fit Statistics

  -2 Res Log Likelihood           818.9
  AIC (smaller is better)         822.9
  AICC (smaller is better)        823.0
  BIC (smaller is better)         825.5
```

The unstructured model has the best (smallest) AIC value (767.3) and is thus to be preferred.

(2) I use the Kenward-Roger method to approximate the denominator degrees of freedom.

```
/*AR(1) model*/
proc mixed data=c9;
class week rat trt;
model weight=week trt week*trt/DDFM=KR;
repeated week/subject=rat*trt type=AR(1);
run;
```

```
         Type 3 Tests of Fixed Effects

                Num     Den
 Effect          DF      DF    F Value    Pr > F

 week             4    94.5     521.58    <.0001
 trt              2    24.6       5.83    0.0085
 week*trt         8    94.8      13.00    <.0001
```

There is a significant interaction ($p < 0.0001$).

(3) The worst fitting model from (1) is the CS model.

In order to test the lack of fit, I am duplicating the time variable into a variable LACKFIT. Moreover, I am duplicating the time variable into a variable T, which will be used for regression. I still need the time variable itself to define a subject effect with the REPEATED statement. I am using the HTYPE=1 option on order to produce “sequential F-tests”.

```
data c9;
set c9;
t=week;
lackfit=week;

proc mixed data=c9;
class week rat trt lackfit;
model weight=trt t trt*t trt*lackfit/DDFM=KR htype=1;
repeated week/subject=rat*trt type=CS;
run;
```

```
         Type 1 Tests of Fixed Effects

                Num     Den
Effect           DF      DF    F Value    Pr > F

trt               2      24       7.63    0.0027
t                 1      96    2863.70    <.0001
t*trt             2      96      59.54    <.0001
trt*lackfit       9      96       1.42    0.1914
```

There is no significant lack of fit (p=0.1914). Now the same analysis using the unstructured model:

```
proc mixed data=c9;
class week rat trt lackfit;
model weight=trt t trt*t trt*lackfit/DDFM=KR htype=1;
repeated week/subject=rat*trt type=UN;
run;
```

```
        Type 1 Tests of Fixed Effects

                Num     Den
Effect           DF      DF    F Value    Pr > F

trt               2      24       7.63    0.0027
t                 1      24     865.39    <.0001
t*trt             2      24      17.99    <.0001
trt*lackfit       9    31.5       6.04    <.0001
```

The lack of fit now is highly significant ($p < 0.0001$)!

Exercise 25

```
/*basic analysis - all genotypes fixed*/
proc mixed data=a5 lognote;
class entry_nr rep block;
model yield=entry_nr rep;
random block*rep;
run;
```

For hybrids, we can fit a two-factorial random-effects model with factors `parent1` and `parent2`. The challenge is that we do not want to fit these effects for the five standards. To achieve this, we fit a separate effect for each standard, plus one single fixed effect to model the expected value of all hybrids. This is done using the factor `standard`. To prevent random effects from being fitted for standards, we define a dummy variable `w` with `w=0` for standards and `w=1` for hybrids. Then we multiply all random hybrid effects with this dummy variable.

```
/*hybrids random, standards fixed*/
data a5;
set a5;
if entry_nr in (1,2,3,4,5) then w=0; else w=1;

proc mixed data=a5 lognote;
class parent1 parent2 standard rep block;
model yield=standard rep;
random w*parent1 w*parent2 w*parent1*parent2 block*rep;
run;

/*hybrids random, standards fixed, with covariate GPF*/
proc mixed data=a5 lognote;
class parent1 parent2 standard rep block;
model yield=standard rep GPF;
random w*parent1 w*parent2 w*parent1*parent2 block*rep/solution;
run;
```

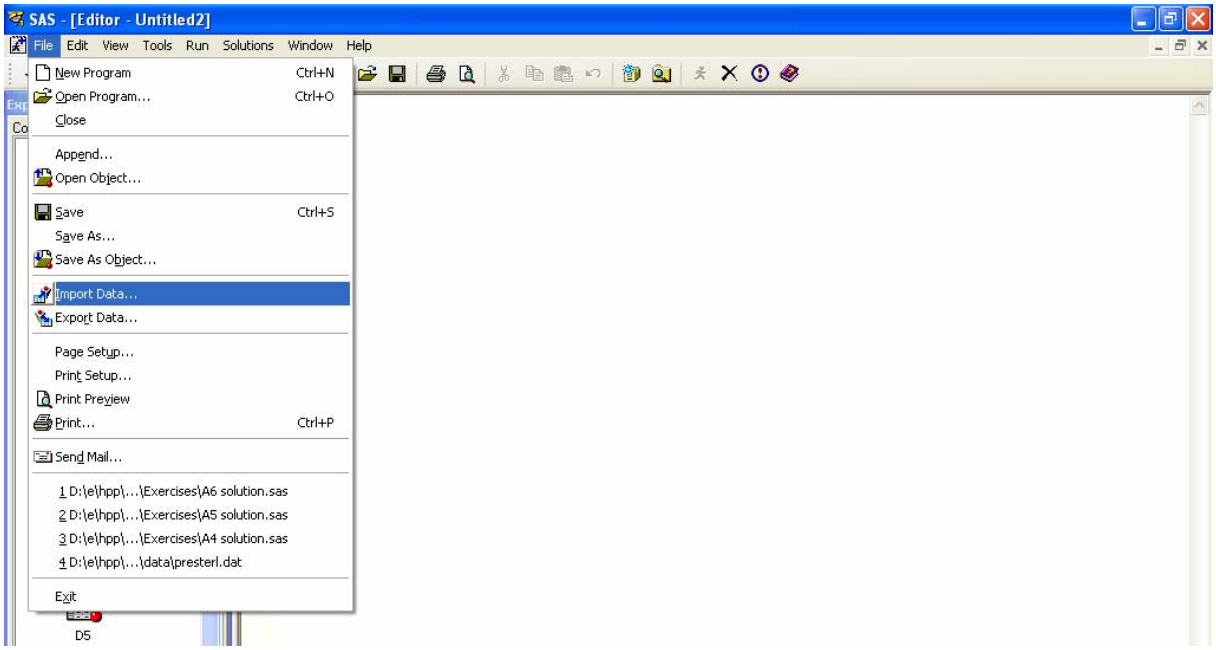
It is tedious to compose estimates of genotypic values from the BLUPs of effect for **parent1**, **parent2** and **parent1*parent2**. Here is some code that is given without further explanation (but see annotation to SAS code).

```
/*hybrids random, standards fixed*/
data a5;
set a5;
if entry_nr in (1,2,3,4,5) then w=0; else w=1;

proc mixed data=a5 lognote;
ods output solutionR=BLUP;
class parent1 parent2 standard rep block;
model yield=standard rep/solution;
random w*parent1 w*parent2 w*parent1*parent2 block*rep/solution;
run;

/*compose genotypic values from parent1, parent2 and cross effect
(parent1*parent2)*/
proc sort data=a5 out=a5;
by entry_nr;
run;

/*get parent1, parent2 and entry_nr codes for genotypes*/
proc means data=a5 noprint;
by entry_nr;
id parent1 parent2;
output out=b5 mean=;
var yield;
run;

/*get design matrix for genotypic values*/
proc glmmod data=b5 outdesign=c5;
class parent1 parent2;
model yield=parent1 parent2 parent1*parent2/noint;
run;

/*get BLUPs of BLUPS for parent1, parent2 and parent1*parent2*/
data BLUP;
set BLUP;
if effect ne 'rep*block';
keep estimate;
run;

/*put it all together in IML and compute K*u, where u is a vector of
  BLUPs for random effects and K is a design matrix for genotypic values*/
proc iml;
use c5;
read all var _num_ into K; /*K is design matrix for genotypic effects*/
close c5;
nc=ncol(K);
K=K[,2:nc]; /*get rid of first column*/
use BLUP;
read all var _NUM_ into u;
close BLUP;
genotypic_value=K*u;
print genotypic_value;
create gvalue var {genotypic_value};
append;
quit;
run;
```

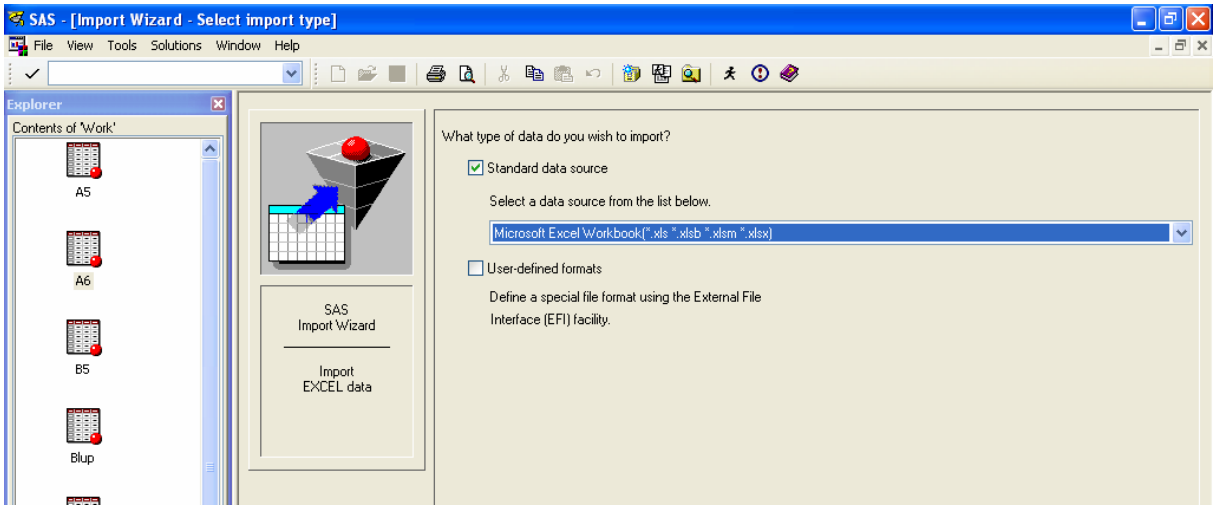
```
/*merge BLUPs of genotypic values with codes for entry_nr, parent1, and
parent2*/
data d5;
merge b5 gvalue;
keep parent1 parent2 entry_nr genotypic_value;
run;

proc print data=d5; run;.
```

Exercise 26

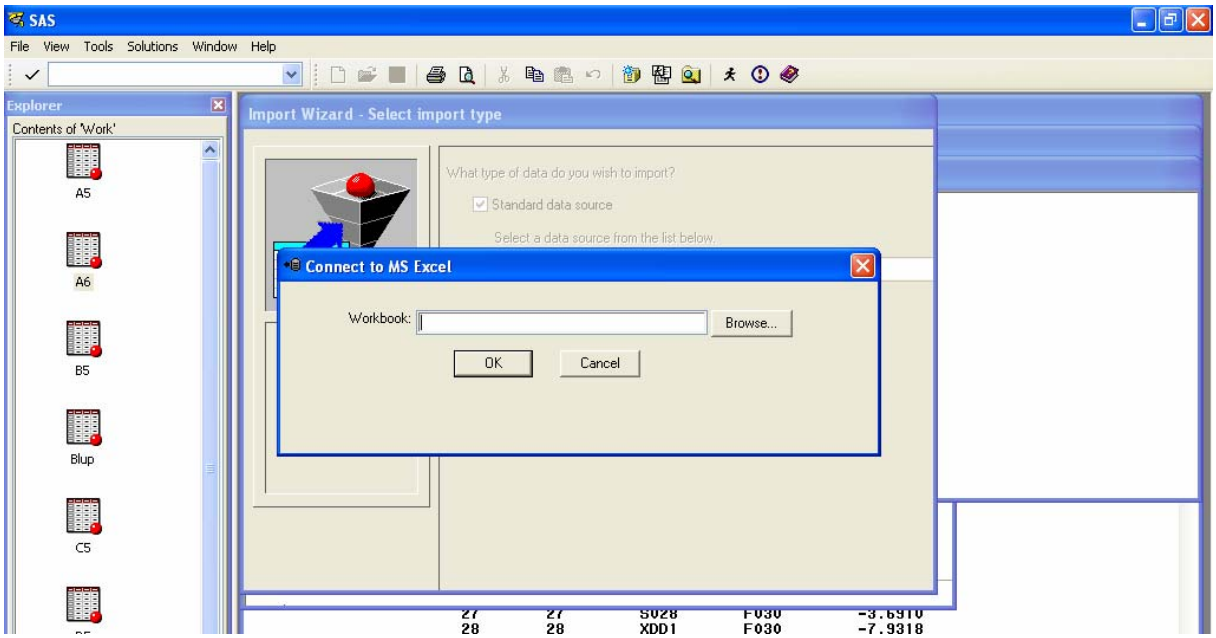
You can use the import facility to import the data from triticale.xls.



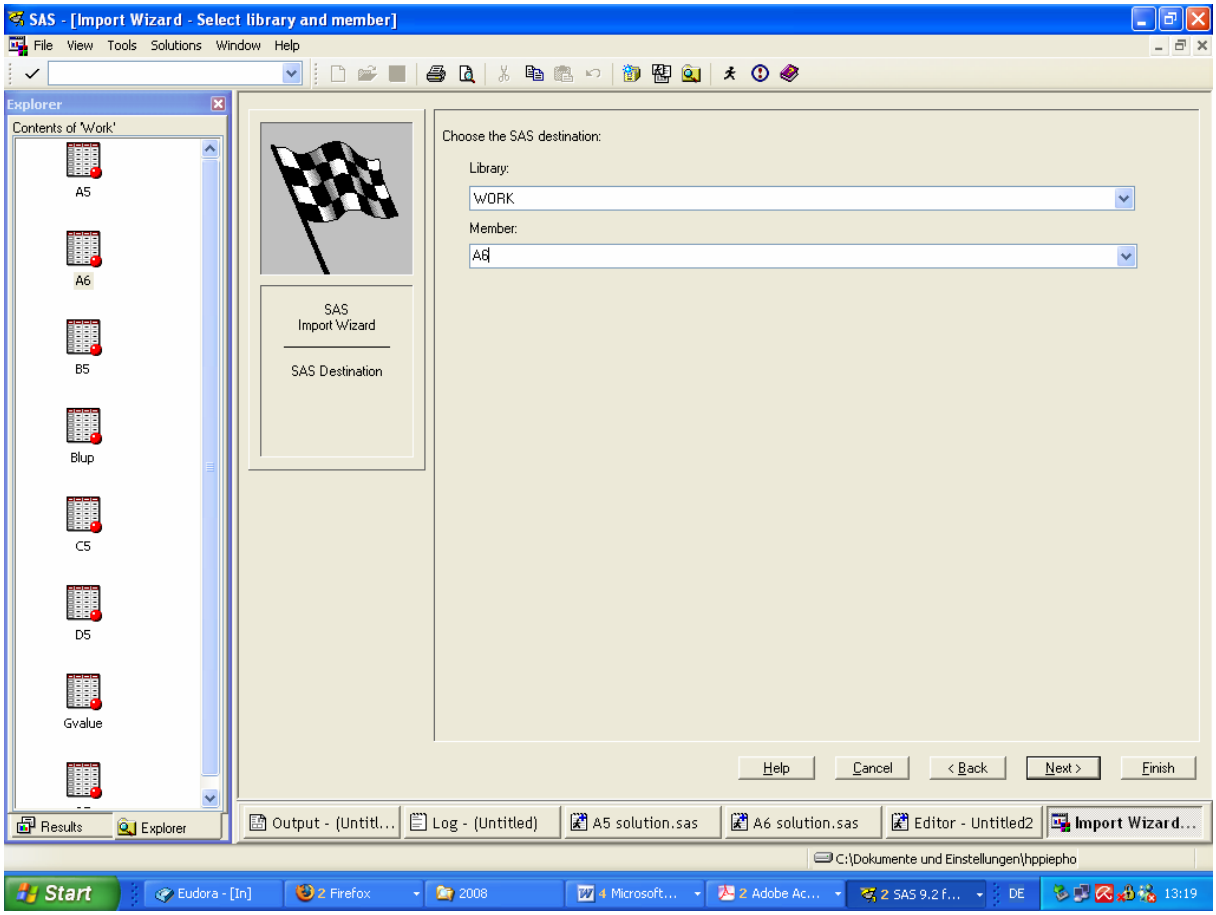
The next you see after clicking on “import data” is this:



The default file types include *.xls files, so simply click on “next”.



Browse to find your Excel-file. Once the path is includes, press OK. Next, you are asked to select the sheet and to assign a SAS filename, under which the imported file is to be stored. Here I select A6 as a SAS filename:



Text, you are given the options to either “Finish”, which finalizes the import, or the menu will also give you the option to save the PROC IMPORT code generated for this task if you click on “Next”. You then need to “browse” again to select the destination and name of SAS program file to which you want to save the code. The code generated here is as follows:

```
PROC IMPORT OUT= WORK.A6
            DATAFILE= "D:\e\hpp\Module\Bioinformatics\data\triticale.xls"
            DBMS=EXCEL REPLACE;
     RANGE="data";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
```

Thus, SAS uses a procedure called IMPORT for this task.

This basic model fits the following **block model** (design effects), expressed using the syntax described in Piepho et al. (2003):

```
location/replicate/block =

location +
location*replicate +
location*replicate*block
```

All effects, including the location main effect are random, if the factor location is considered as random. This could be specified in a RANDOM statement as follows:

```
random location location*replicate replicate*block;
```

But this is computationally demanding. It helps to define locations as independent "subjects" to speed up the convergence (see Chapter on repeated measures and Piepho & Möhring, 2011). The code is

```
random int replicate replicate*block/sub=location;
```

The **treatment model** can be built using factors check, lines, hybrids and seed_density. The classification is hierarchical with check at the top of the hierarchy. Lines as well as hybrids are nested within check. Also, lines must be crossed with lines because there are two seed densities with lines. The model can be written

```
check/(lines*seed_density + hybrids) =

check +
check*lines*seed_density +
check*hybrids
```

Finally, it is a good idea to allow for interaction between treatment effects and the factor location. This interaction is random, because the factor location is random. Essentially, all effects of the treatment model need to be crossed with the factor location, and this is done most effectively using location as a subject-effect as follows:

```
random check check*lines*seed_density check*hybrids/sub=location;
```

A small modification is needed here, however, because variance components for the nested effects involving **`check*lines*seed_density` and `check*hybrids`** is to be fitted only

for lines and hybrids, respectively, but not for checks. This can be achieved by using the two dummy variables **dummy_line** and **dummy_hybrid** as follows:

```
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids/sub=location;
```

The full code to fit this model is this:

```
/*basic model, genotypes fixed*/
proc mixed data=a6 lognote;
class location replicate block check lines hybrids seed_density;
model yield=check check*lines*seed_density check*hybrids;
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids/sub=location;
       /*genotype-environment interaction effects*/
random int replicate replicate*block/sub=location;
       /*design effects*/
run;
```

```
      Covariance Parameter Estimates

Cov Parm                 Subject     Estimate

Check                    location      2.1956
dumm*Chec*line*seed_     location     19.7299
dummy_*Check*hybrids     location     15.4989
Intercept                location      156.83
replicate                location     0.09590
replicate*block          location     18.0131
Residual                              21.1007


              Fit Statistics

    -2 Res Log Likelihood         17062.7
    AIC (smaller is better)       17076.7
    AICC (smaller is better)      17076.8
    BIC (smaller is better)       17074.0
```

Next, we consider fitting random effects for lines and hybrids, but not for checks. This is done using the same two dummy variables. Be ware, however, that computing time is considerable prolonged for this code, because there no longer is a common subject effect to all RANDOM statements (see Piepho & Möhring, 2011), meaning that the model no longer has independent subjects in the repeated measures terminology.

```
/*basic model, genotypes random*/
proc mixed data=a6 lognote;
class location replicate block check lines hybrids seed_density;
model yield=;
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids;
       /*genotypic main effects*/
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids/sub=location;
       /*genotype-environment interaction effects*/
random int replicate replicate*block/sub=location;
       /*design effects*/
run;
```

```
                    Convergence criteria met.


                  Covariance Parameter Estimates

Cov Parm                  Subject     Estimate

Check                                  29.7768
dumm*Chec*line*seed_                   18.8895
dummy_*Check*hybrids                   16.2935
Check                     location      2.1729
dumm*Chec*line*seed_      location     19.6945
dummy_*Check*hybrids      location     15.4690
Intercept                 location      156.82
replicate                 location      0.1101
replicate*block           location     17.5788
Residual                               21.1571

                      The Mixed Procedure

                        Fit Statistics

    -2 Res Log Likelihood          18601.6
    AIC (smaller is better)        18621.6
    AICC (smaller is better)       18621.6
    BIC (smaller is better)        18619.5
```

Next, we fit a g.c.a. effects to model the hybrid main effect. I this is added, to the previous model, the simple main effect for hybrids becomes an s.c.a. effect. The following code can be used to generate 21 dummy variables for g.c.a. effects, when the dataset a6 contains variables **`parent1`** and **`parent2`** for the two parents:

```
data b6;
set a6;
array gca gca1-gca21 (21*0);
do i = 1 to 21;
  gca(i) = (parent1=i) or (parent2=i);
end;
drop i;
run;
```

This code assigns a value of 0 to all 21 dummy variables gca1-gca21.

The following random statement is added to the previous code to add a g.c.a. effect for hybrids:

```
random dummy_hybrid*gca1
       dummy_hybrid*gca2
       dummy_hybrid*gca3
       dummy_hybrid*gca4
       dummy_hybrid*gca5
       dummy_hybrid*gca6
       dummy_hybrid*gca7
       dummy_hybrid*gca8
       dummy_hybrid*gca9
       dummy_hybrid*gca10
       dummy_hybrid*gca11
       dummy_hybrid*gca12
       dummy_hybrid*gca13
       dummy_hybrid*gca14
       dummy_hybrid*gca15
       dummy_hybrid*gca16
       dummy_hybrid*gca17
```

```
        dummy_hybrid*gca18
        dummy_hybrid*gca19
        dummy_hybrid*gca20
        dummy_hybrid*gca21 / type=toep(1);
```

The option `type=toep(1)` selects a so-called Toeplitz covariance structure, which is needed here to ensure that all 21 g.c.a. effects have the same variance (check online documentation to read more about this covariance structure; also see Möhring et al., 2011). The full code is as follows:

```
/*basic model, genotypes random, g.c.a. effect for hybrids*/
proc mixed data=b6 lognote;
class location replicate block check lines hybrids seed_density;
model yield=;
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids;
        /*genotypic main effects*/
random dummy_hybrid*gca1
        dummy_hybrid*gca2
        dummy_hybrid*gca3
        dummy_hybrid*gca4
        dummy_hybrid*gca5
        dummy_hybrid*gca6
        dummy_hybrid*gca7
        dummy_hybrid*gca8
        dummy_hybrid*gca9
        dummy_hybrid*gca10
        dummy_hybrid*gca11
        dummy_hybrid*gca12
        dummy_hybrid*gca13
        dummy_hybrid*gca14
        dummy_hybrid*gca15
        dummy_hybrid*gca16
        dummy_hybrid*gca17
        dummy_hybrid*gca18
        dummy_hybrid*gca19
        dummy_hybrid*gca20
        dummy_hybrid*gca21 / type=toep(1);
          /*g.c.a. effect for hybrids*/
random check dummy_line*check*lines*seed_density
dummy_hybrid*check*hybrids/sub=location;
        /*genotype-environment interaction effects*/
random int replicate replicate*block/sub=location;
        /*design effects*/
run;
```

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	
Check		31.5303	
dumm*Chec*line*seed_		18.8971	
dummy_*Check*hybrids		16.1068	
Variance		**1.7653**	**-> g.c.a. variance**
Check	location	1.9912	
dumm*Chec*line*seed_	location	19.6617	
dummy_*Check*hybrids	location	15.3466	

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	location	156.50
replicate	location	0.1084
replicate*block	location	17.6211
Residual		21.1688

Fit Statistics

-2 Res Log Likelihood	18600.7
AIC (smaller is better)	18622.7
AICC (smaller is better)	18622.8
BIC (smaller is better)	18620.4

The AIC is slightly increased compared the previous model, indicating that fitting g.c.a. effects did not improve the fit here.

Exercise 36

```
data rice;
input
    block    n    v    n_amount    yield;
datalines;
      1      1    1           0       4520
      1      1    2           0       4034
      1      1    3           0       3554
      1      1    4           0       4216
      1      2    1          60       5598
      1      2    2          60       6682
      1      2    3          60       4948
      1      2    4          60       5372
      1      3    1          90       5806
      1      3    2          90       5738
      1      3    3          90       5974
      1      3    4          90       4276
      1      4    1         120       6192
      1      4    2         120       6869
      1      4    3         120       5522
      1      4    4         120       2504
      1      5    1         150       7470
      1      5    2         150       7862
      1      5    3         150       7260
      1      5    4         150       1594
      1      6    1         180       8542
      1      6    2         180       6318
      1      6    3         180       5684
      1      6    4         180       2338
      2      1    1           0       4208
      2      1    2           0       5044
      2      1    3           0       2674
      2      1    4           0       4212
      2      2    1          60       5256
      2      2    2          60       5948
      2      2    3          60       6094
      2      2    4          60       4694
```

```
      2       3     1         90        6600
      2       3     2         90        6307
      2       3     3         90        5904
      2       3     4         90        5924
      2       4     1        120        7146
      2       4     2        120        7072
      2       4     3        120        5970
      2       4     4        120        5126
      2       5     1        150        7578
      2       5     2        150        6324
      2       5     3        150        6392
      2       5     4        150        1690
      2       6     1        180        9012
      2       6     2        180        7567
      2       6     3        180        7302
      2       6     4        180        1560
      3       1     1          0        4030
      3       1     2          0        3840
      3       1     3          0        3304
      3       1     4          0        5016
      3       2     1         60        6162
      3       2     2         60        5316
      3       2     3         60        5286
      3       2     4         60        4382
      3       3     1         90        6794
      3       3     2         90        6732
      3       3     3         90        6104
      3       3     4         90        4236
      3       4     1        120        6860
      3       4     2        120        6744
      3       4     3        120        6550
      3       4     4        120        3818
      3       5     1        150        7642
      3       5     2        150        6666
      3       5     3        150        6410
      3       5     4        150        2856
      3       6     1        180        8548
      3       6     2        180        5736
      3       6     3        180        5210
      3       6     4        180        1744
;
/*check if quadratic term is significant for v=1*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
run;

/*define variable switch*/
data rice;
set rice;
switch=1;
if v=1 then switch=0;
run;

/*fit linear model for variety V1 and quadratic model for V2-V4*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    switch*v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
run;
```

```
/*test of coincidence for varieties V2 and V3*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
contrast 'coincidence of V=2 and V=3' v 0 1 -1 0,
                                      v*n_amount 0 1 -1 0,
                                       v*n_amount*n_amount 0 1 -1 0;
run;

/*test of coincidence for all four varieties*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
contrast 'coincidence of all 4 var.'  v 1 -1 0 0,
                                      v 1 0 -1 0,
                                      v 1 0 0 -1,
                                      v*n_amount 1 -1 0 0,
                                      v*n_amount 1 0 -1 0,
                                      v*n_amount 1 0 0 -1,
                                       v*n_amount*n_amount 1 -1 0 0,
                                       v*n_amount*n_amount 1 0 -1 0,
                                       v*n_amount*n_amount 1 0 0 -1;
run;

/*test of coincidence for all four varieties - a 2nd solution*/
proc mixed data=rice;
class block n v;
model yield=block v v*n_amount
                    v*n_amount*n_amount/ ddfm=KR solution;
random n*block; /*main plot error*/
contrast 'coincidence of all 4 var.'  v 1 -1 0 0,
                                      v 1 1 -2 0,
                                      v 1 1 1 -3,
                                      v*n_amount 1 -1 0 0,
                                      v*n_amount 1 1 -2 0,
                                      v*n_amount 1 1 1 -3,
                                       v*n_amount*n_amount 1 -1 0 0,
                                       v*n_amount*n_amount 1 1 -2 0,
                                       v*n_amount*n_amount 1 1 1 -3;
run;
```