



Reaml

# **A Guide to REML in Genstat® (21<sup>st</sup> Edition)**

by Roger Payne, Sue Welham and Simon Harding.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,  
Hemel Hempstead, Hertfordshire HP2 4TP, UK

E-mail: [info@genstat.co.uk](mailto:info@genstat.co.uk)

Website: <http://www.genstat.co.uk/>

First published 2008, for GenStat *for Windows* 11<sup>th</sup> Edition

This edition published 2020, for Genstat *for Windows* 21<sup>st</sup> Edition

Genstat is a registered trade of **VSN International**. All rights reserved.

© 2020 VSN International

---

# Contents

## Introduction [1](#)

### 1 Linear mixed models [2](#)

- 1.1 Split-plot design [3](#)
- 1.2 Commands for REML analysis [19](#)
- 1.3 Practical [25](#)
- 1.4 Means plots [26](#)
- 1.5 Practical [30](#)
- 1.6 Predictions [30](#)
- 1.7 Practical [38](#)
- 1.8 A non-orthogonal design [38](#)
- 1.9 Practical [45](#)
- 1.10 Residual checks and plots [46](#)
- 1.11 Practical [49](#)
- 1.12 Saving information from the analysis [49](#)

### 2 Meta analysis with REML [51](#)

- 2.1 Example: a series of fungicide trials [52](#)
- 2.2 Commands for meta analysis [59](#)
- 2.3 Practical [61](#)

### 3 Spatial analysis [62](#)

- 3.1 Traditional blocking [63](#)
- 3.2 Correlation modelling [65](#)
- 3.3 The VSTRUCTURE directive [78](#)
- 3.4 Practical [84](#)
- 3.5 The variogram [84](#)
- 3.6 Practical [86](#)
- 3.7 Determining the random and correlation models automatically [86](#)
- 3.8 Practical [92](#)
- 3.9 Automatic analysis of series of trials [93](#)
- 3.10 Practical [99](#)

### 4 Repeated measurements [100](#)

- 4.1 Correlation models over time [101](#)
- 4.2 Practical [107](#)
- 4.3 Random coefficient regression [107](#)
- 4.4 Practical [112](#)

## Index [113](#)



---

## Introduction

The REML algorithm provides several important types of analysis, that are useful in a wide range of application areas including biology, medicine, industry and finance. In biology they are usually known as *linear mixed models*, but in some application areas (e.g. education) they may be called *multi-level models*.

Genstat's REML facilities are powerful and comprehensive, but nevertheless very straightforward and easy to use. This book is designed to introduce you to these techniques, and give you the knowledge and confidence to use them correctly and effectively. It has been written to provide the notes for VSN's course on the use of REML in Genstat, but it can be used equally well as a self-learning tool.

One of the key features of REML is that it can analyse data that involve more than one source of error variation. In this respect it is similar to the Genstat ANOVA algorithm, and the similarities and differences between the two methods are explored in detail in Chapter 1. An important advantage of REML over ANOVA is that it can analyse unbalanced designs. It also has a powerful prediction algorithm that extends the ideas in Genstat's regression prediction algorithm to cover random as well as fixed effects.

Chapter 2 covers the use of REML for meta analysis, showing how you can do a simultaneous analysis of several disparate data sets to obtain combined estimates for the treatments of interest.

A further advantage of REML is explored in Chapter 3, where we show how it can model spatial correlations between observations in two-dimensions. These methods have proved very successful, for example in the analysis of field experiments to assess new plant varieties. The designs often contain too many varieties for the conventional blocking techniques (e.g. the use of randomized-block designs) to be effective. So instead, for example, auto-regressive models are fitted to the spatial correlations across the field.

Chapter 4 examines the use of correlation modelling in the analysis of repeated measurements. Here the correlation is in a single dimension, namely time, and REML provides a powerful alternative to conventional methods such as repeated-measures ANOVA or the analysis of contrasts over time.

The book works through a series of straightforward examples, with frequent practicals to allow you to try the methods for yourself. The examples work mainly through the menus of Genstat *for Windows*, so there is no need for prior knowledge of the Genstat command language. However, we do assume that you will be familiar with ordinary analysis of variance. (If not, we recommend that you work through Chapters 1-5 of the *Guide to ANOVA and Design in Genstat*.)

---

# 1 Linear mixed models

The **REML** algorithm is designed to analyse *linear mixed models* (also known as *multi-level models*). The word *mixed* here indicates that the model contains *fixed* terms like treatments, as well as *random* terms, like rows and columns of a field experiment or aspects such as litters in animal experiments. The important feature of **REML** is that it can handle several random terms (in addition to the usual residual term). The Genstat **ANOVA** algorithm can also handle several random terms, and we start by comparing the analyses from **ANOVA** with those from **REML**.

In this chapter you will learn

- how to use the **Linear Mixed Models** (and **Analysis of Variance** ★) menus
- what output is given by a Genstat **REML** analysis, and how it compares to Genstat **ANOVA**
- how to assess treatment terms by Wald and F statistics
- how to plot means
- how to form predictions
- how to plot (and assess) residuals
- the commands **VCOMPONENTS**, **REML**, **VDISPLAY**, **VGRAPH**, **VPREDICT** and **VPLOT** ★

Note: the topics marked ★ are optional.

## 1.1 Split-plot design

The design used most often to illustrate the need for several random (or error) terms in ANOVA is the split-plot design.

V3 N3	V3 N2	V3 N2	V3 N3
V3 N1	V3 N0	V3 N0	V3 N1
V1 N0	V1 N1	V2 N0	V2 N2
V1 N3	V1 N2	V2 N3	V2 N1
V2 N0	V2 N1	V1 N1	V1 N2
V2 N2	V2 N3	V1 N3	V1 N0
V3 N2	V3 N0	V2 N3	V2 N0
V3 N1	V3 N3	V2 N2	V2 N1
V1 N3	V1 N0	V1 N2	V1 N3
V1 N1	V1 N2	V1 N0	V1 N1
V2 N1	V2 N0	V3 N2	V3 N3
V2 N2	V2 N3	V3 N1	V3 N0
V2 N1	V2 N2	V1 N2	V1 N0
V2 N3	V2 N0	V1 N3	V1 N1
V3 N3	V3 N1	V2 N3	V2 N2
V3 N2	V3 N0	V2 N0	V2 N1
V1 N0	V1 N3	V3 N0	V3 N1
V1 N1	V1 N2	V3 N2	V3 N3

In the *split-plot* design shown here, the treatments are three varieties of oats (*Victory*, *Golden rain* and *Marvellous*) and four levels of nitrogen (0, 0.2, 0.4 and 0.6 cwt). As it is feasible to work with smaller plots for fertiliser than for varieties, the six blocks were initially split into three whole-plots and then each whole-plot was split into four subplots. The varieties were allocated (at random) to the whole-plots within each block, and the nitrogen levels (at random) to the subplots within each whole-plot. In a randomized-block design, we have a hierarchical structure with blocks and then plots within blocks.

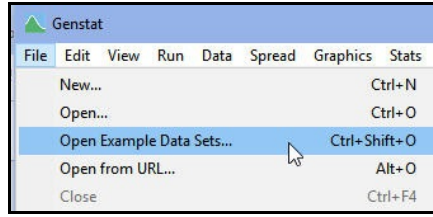


Figure 1.1

The data files for the examples and exercises used in this Guide can be accessed using the [Example Data Sets](#) menu (Figure 1.2). Click on [File](#) on the menu bar, and select the [Open Example Data Sets](#) option, as shown in Figure 1.1. In the menu, it is convenient to "filter" by the topic [A Guide to REML](#) using the drop-down list box in the upper part of the menu. The menu will then list only the files used in this Guide. The data for the split-plot experiment are in the file [Oats.gsh](#).

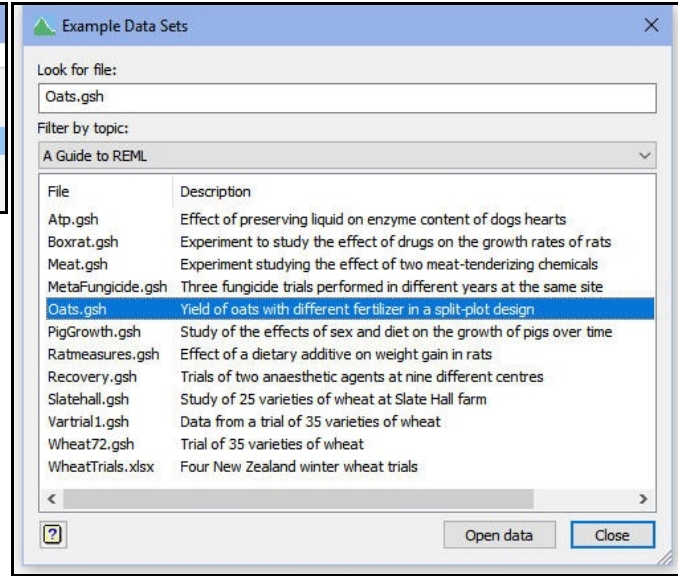


Figure 1.2

The model describes the yield  $y_{ijk}$  from block  $i$ , whole-plot  $j$ , subplot  $k$  by the equation

$$y_{ijk} = \mu + v_r + a_s + va_{rs} + b_i + w_{ij} + \varepsilon_{ijk}$$

where the *fixed* part of the model consists of

- $\mu$  the overall constant (grand mean),
- $v_r$  the main effect of variety  $r$  (where  $r$  is the variety assigned to unit  $ijk$ ),
- $a_s$  the main effect of nitrogen application at level  $s$  (where  $s$  is the nitrogen level assigned to subplot  $ijk$ ), and
- $va_{rs}$  their interaction.

The *random* model terms are

- $b_i$  the effect of block  $i$ ,
- $w_{ij}$  the effect of whole-plot  $j$  within block  $i$ , and
- $\varepsilon_{ijk}$  the random error (i.e. residual) for unit  $ijk$  (which here is the same as the subplot effect, since the subplots are the smallest units of the experiment).

The model can be written in matrix notation as

$$y = \sum_i \mathbf{X}_i \beta_i + \sum_i \mathbf{Z}_i u_i + \varepsilon$$

where

- $y$  is the vector of data values,
- $\beta_i$  is the vector of fixed effects for treatment term  $i$  with design matrix  $\mathbf{X}_i$ ,
- $u_i$  is the vector of random effects for random term  $i$  with design matrix  $\mathbf{Z}_i$ ,
- $\varepsilon$  is the vector of residuals.



To analyse the split-plot by ANOVA, you select the **General** sub-option of the **Analysis of Variance** option of the **Stats** menu on the menu bar. You can customize the menu for the split plot by selecting **Split-plot design** in the **Design** drop-down list box in the menu (see Figure 1.3).

For this data set, the treatment structure is a factorial with the

two factors, **nitrogen** and **variety**. The block structure is set up automatically by Genstat from the factors specified in the **Blocks**, **Whole plots** and **Sub-plots** fields.

The analysis-of-variance table shows that we have three *strata* in the hierarchy, corresponding to the three random terms: blocks, whole-plots within blocks, and subplots within whole plots (within blocks). The analysis automatically works out where each fixed (or treatment) term is estimated, and compares it with the correct residual. So the sum of squares for **variety** (which was applied to complete whole-plots) is compared with a residual which represents the random variability of the whole-plots. Conversely, **nitrogen** (which was applied to subplots) and the **variety.nitrogen** interaction are compared with the residual for subplots within whole-plots.

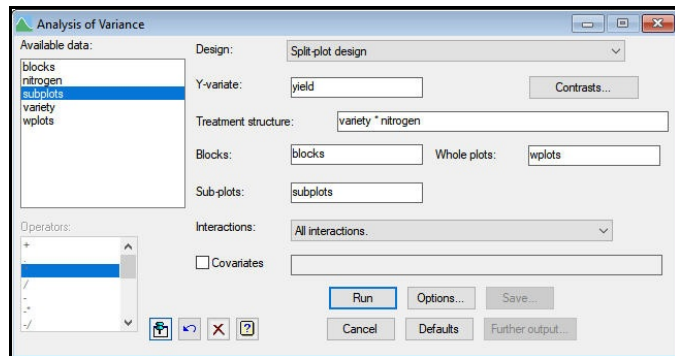


Figure 1.3

## Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
blocks stratum	5	15875.3	3175.1	5.28	
blocks.wplots stratum					
variety	2	1786.4	893.2	1.49	0.272
Residual	10	6013.3	601.3	3.40	
blocks.wplots.subplots stratum					
nitrogen	3	20020.5	6673.5	37.69	<.001
variety.nitrogen	6	321.8	53.6	0.30	0.932
Residual	45	7968.8	177.1		
Total	71	51985.9			

*Message: the following units have large residuals.*

blocks 1	31.4	s.e. 14.8
----------	------	-----------

## Tables of means

Variate: yield

Grand mean 104.0

variety	Victory 97.6	Golden rain 104.5	Marvellous 109.8		
nitrogen	0 cwt 79.4	0.2 cwt 98.9	0.4 cwt 114.2	0.6 cwt 123.4	
variety	nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
Victory		71.5	89.7	110.8	118.5
Golden rain		80.0	98.5	114.7	124.8
Marvellous		86.7	108.5	117.2	126.8

## Standard errors of differences of means

Table	variety	nitrogen	variety nitrogen
rep.	24	18	6
s.e.d.	7.08	4.44	9.72
d.f.	10	45	30.23
Except when comparing means with the same level(s) of variety			7.68
d.f.			45

The standard errors accompanying the tables of means also take account of the stratum where each treatment term was estimated. The `variety` s.e.d. of

$$7.08 = \sqrt{(2 \times 601.3 / 24)}$$

is based on the residual mean square for `blocks.wplots`, while that for `nitrogen`

$$4.44 = \sqrt{(2 \times 177.1 / 18)}$$

is based on that for `blocks.wplots.subplots`. The `variety × nitrogen` table is more interesting. There are two s.e.d.'s according to whether the two means to be compared are for the same variety. If they are, then the subplots from which the means are calculated will all involve the same set of whole-plots, so any whole-plot variability will cancel out, giving a smaller s.e.d. than for a pair of means involving different varieties.

Split-plot designs do not occur only in field experiments, but they can occur in animal trials (where, for example, the same diet may need to be fed to all the animals in a pen but other treatments may be applied to individual animals), or in industrial experiments (where different processes may require different sized batches of material), or even in cookery experiments (see, for example, Cochran & Cox 1957, page 299). There can also be more than one treatment factor applied to the units of any stratum. To analyse the results in Genstat, you simply need to specify the blocking factors, as above, and then whatever treatment structure is appropriate.

Genstat specifies the structure of the design, and thus the different sources of variability (or strata) in the model, using the `BLOCKSTRUCTURE` directive. For Figure 1.3, this was defined automatically as

```
BLOCKSTRUCTURE blocks / wplots / subplots
```

here the operator `/` indicates that a factor is nested within another factor. So we have `subplots` nested within `wplots` (whole-plots) nested within `blocks`, as required. The model formula expands to the list of model terms

```
blocks + blocks.wplots + blocks.wplots.subplots
```

which defines the strata to represent the variation between the blocks, between whole-plots within blocks, and between subplots within whole plots (within blocks) shown in the analysis-of-variance table.

The treatment (or fixed) terms to be fitted in the analysis are specified by the `TREATMENTSTRUCTURE` directive. The menu in Figure 1.3 uses the setting of the `Treatment structure` box to define these as

```
TREATMENTSTRUCTURE variety * nitrogen
```

which expands to give the model terms

```
variety + nitrogen + variety.nitrogen
```

The analysis of variance is done by the `ANOVA` directive:

```
ANOVA yield
```

for Figure 1.3. Full details of `BLOCKSTRUCTURE`, `TREATMENTSTRUCTURE` and `ANOVA` are given in the *Guide to the Genstat Command Language, Part 2 Statistics*, Chapter 4, and further information about the Analysis of Variance menus is in the *Guide to Anova and Design in Genstat*.

The Analysis of Variance menus deal mainly with balanced designs. This ideal situation, however, is not always achievable. The split-plot design here is balanced because every block contains one of each treatment combination, and every whole-plot contains one of each `nitrogen` treatment. However, there may sometimes be so many treatments that the blocks would become unrealistically large. Designs where each block contains less than the full set of treatments include cyclic designs and Alpha designs (both of which can be generated within Genstat by clicking `Stats` on the menu bar, selecting `Design` and then `Select Design`), neither of which tend to be balanced. In experiments on animals, some subjects may fail to complete the experiment for reasons unconnected with the treatments. So even an initially balanced experiment may not yield a balanced set of data for analysis. The `Mixed Models (REML)` menus, which use the Genstat `REML` directive, are designed to handle these situations.

To analyse the design using REML, select the **Linear Mixed Models** sub-option of the **Mixed Models (REML)** option of the **Stats** menu on the menu bar.

The **Fixed model** box in the menu (Figure 1.4) corresponds to the **Treatment structure** box in the split-plot menu, and specifies the terms defining the *fixed* effects in the model to be fitted. The **Linear Mixed Models** menu provides general facilities

covering any type of design, and so the *random* effects are defined explicitly by the contents of the **Random model** box, instead of being defined automatically as in the split-plot menu. The model is the same though.

The **Options** button produces the menu in Figure 1.5. The standard model options (as shown in the figure) are fine for this design, so we need only select the output to display (and then click **OK**).

Returning to the main menu (Figure 1.4): initial values are seldom required for simple REML analyses like this, and the **Spline model** box is not relevant (this is mainly useful with repeated measurements), so we can click on **Run** and generate the output shown below.

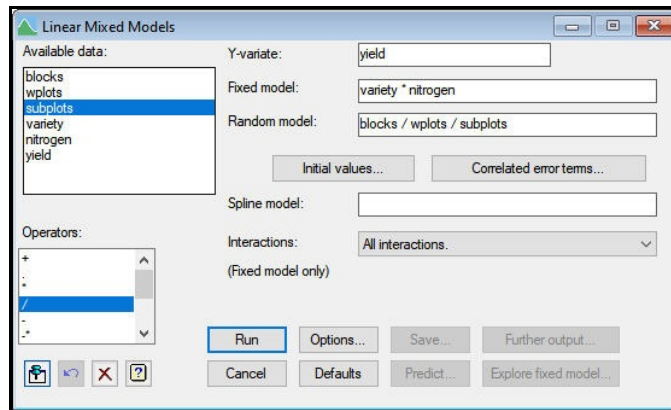


Figure 1.4

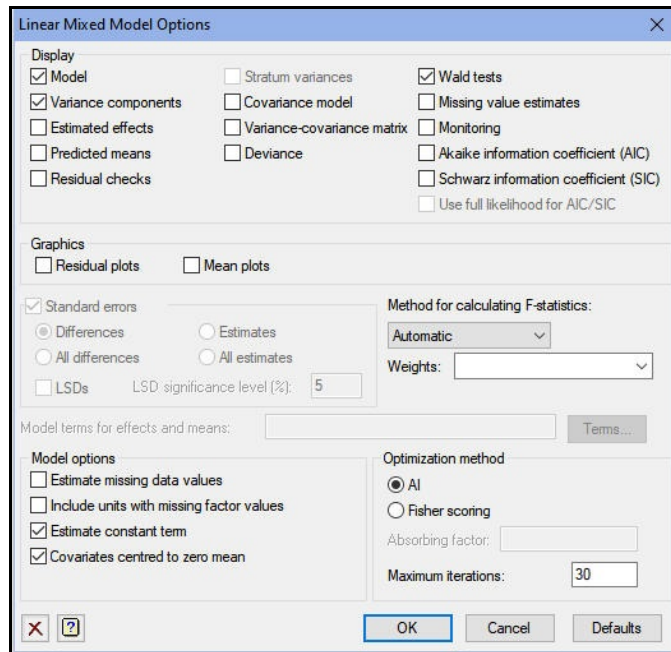


Figure 1.5

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + variety + nitrogen + variety.nitrogen  
 Random model: blocks + blocks.wplots + blocks.wplots.subplots  
 Number of units: 72

blocks.wplots.subplots used as residual term

Sparse algorithm with AI optimisation

## Estimated variance components

Random term	component	s.e.
blocks	214.5	168.8
blocks.wplots	106.1	67.9

## Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
blocks.wplots.subplots	Identity	Sigma2	177.1	37.3

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	2.97	2	1.49	10.0	0.272
nitrogen	113.06	3	37.69	45.0	<0.001
variety.nitrogen	1.82	6	0.30	45.0	0.932

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety.nitrogen	1.82	6	0.30	45.0	0.932

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

The output first lists the terms in the fixed and random model, and indicates the residual term. The residual term is a random term with a parameter for every unit in the design. Here we have specified a suitable term, `blocks.wplots.subplots`, explicitly. However, if we had specified only `blocks + blocks.wplots` as the Random Model (for example by putting `blocks/wplots`), Genstat would have added an extra term `*units*` to act as residual. (`*units*` would be a private factor with a level for every unit in the design.)

Genstat estimates a *variance component* for every term in the random model, apart from the residual. The variance component measures the inherent variability of the term, over and above the variability of the sub-units of which it is composed. Generally, this is positive, indicating that the units become more variable the larger they become. So here the whole-plots are more variable than the subplots, and the blocks are more variable than the whole-plots within the blocks. (This is the same conclusion that you would draw from the analysis-of-variance table earlier in this section and, in fact, you can also produce the variance components as part of the stratum variances output from the [Analysis of Variance](#) menu.) However, the variance component can sometimes be *negative*, indicating that the larger units are *less* variable than you would expect from the contributions of the sub-units of which they are composed. This could happen if the sub-units were negatively correlated.

The section of output summarizing the residual variance model indicates that we have not fitted any specialized correlation model on this term (see the column headed `Model`), and gives an estimate of the residual variance; this is the same figure as is given by the mean square in the residual line in the `blocks.wplots.subplots` stratum in the split-plot analysis-of-variance table.

The next section, however, illustrates a major difference between the two analyses. When the design is balanced, the `Analysis of Variance` menu is able to produce an analysis-of-variance table that partitions the variation into *strata* with an appropriate random error term (or residual) for each treatment term. No such partitioning is feasible for the unbalanced situations that `REML` is designed to handle. Instead Genstat produces a *Wald statistic* to assess each fixed term.

If the design is orthogonal, the Wald statistic is equal to the treatment sum of squares divided by the stratum residual mean square. So under the usual assumption that the residuals come from Normal distributions, the Wald statistic divided by its degrees of freedom will have an F distribution,  $F_{m,n}$ , where  $m$  is the number of degrees of freedom of the fixed term, and  $n$  is the number of residual degrees of freedom for the fixed term. By default, unless the design is large or complicated, Genstat estimates  $n$ , and prints it in the column headed “`d.d.f.`” (i.e. *denominator degrees of freedom*);  $m$  is shown in the column headed “`n.d.f.`” (i.e. *numerator degrees of freedom*). For orthogonal designs, the F statistics and probabilities are identical to those produced by the `Analysis of Variance` menus, and can be used in exactly the same way. In other situations, the printed F statistics have approximate F distributions. So you need to be careful if the value is close to a critical value.

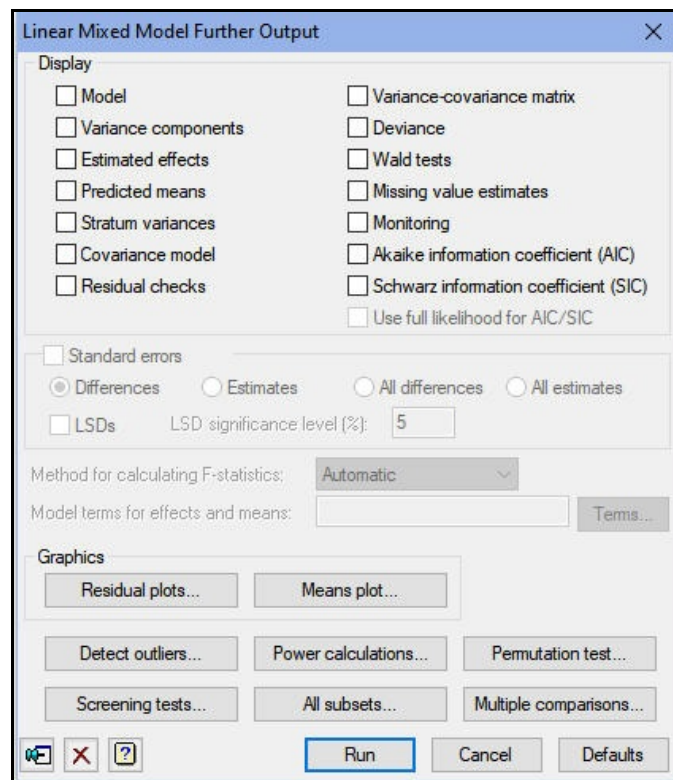
The `Linear Mixed Model Options` menu (Figure 1.5) has a list box `Method for calculating F statistics` to control how and whether to calculate the F statistics. With the default setting, `automatic`, Genstat itself decides whether the statistics can be calculated quickly enough to be useful, and the best method to use. The other settings allow you to select to use either algebraic or numerical derivatives, or to print just Wald statistics (`none`).

The Wald statistics themselves would have exact  $\chi^2$  distributions if the variance parameters were known but, as they must be estimated, they are only asymptotically distributed as  $\chi^2$ . In practical terms, the  $\chi^2$  values will be reliable if the residual degrees of freedom for a fixed term is large compared to its own degrees of freedom. Otherwise they tend to give significant results rather too frequently. The F statistics, if available, are more reliable than the Wald statistics. If they are not calculated, Genstat produces probabilities for the Wald statistics instead, which should again be used with care especially when the value is close to a critical value.

In the example, the treatment terms are *orthogonal* so it makes no difference whether `nitrogen` or `variety` is fitted first. In a non-orthogonal design, however, the ordering of fitting is important, and you should be aware that each test in the “`Sequentially adding terms to fixed model`” section represents the effect of adding the term concerned to a model containing all the terms in the preceding lines. The next section, headed “`Dropping individual terms from full fixed model`” looks at the effect of removing terms from the complete fixed model: so the lines here allow you to assess the effects of a term after eliminating all the other fixed terms. This is particularly useful for seeing how the model might be simplified. Notice that the only relevant term here is the variety by nitrogen interaction. We cannot remove a main effect (such as `nitrogen` or `variety`) from a model that contains an interaction involving that factor.



The **Further output** button generates the menu shown in Figure 1.6, in which we have checked the boxes to produce tables of predicted means with standard errors of differences between means and least significant differences. The **Model terms for effects and means** box enables you to specify the terms for which you want tables of means (and, if you had checked the **Estimated effects** box, tables of effects). The default, which is fine here, is to produce a table for each term in the fixed model. Clicking **Run** then generates the tables shown below. Because the fixed terms are orthogonal, the means are identical to those produced by the **Analysis of Variance** menu.



**Figure 1.6**

### Table of predicted means for Constant

104.0    Standard error: 6.64

### Table of predicted means for variety

variety	Victory	Golden rain	Marvellous
	97.6	104.5	109.8

Standard error of differences: 7.079

### Table of predicted means for nitrogen

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	79.4	98.9	114.2	123.4

Standard error of differences: 4.436

## Table of predicted means for variety.nitrogen

nitrogen variety	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
Victory	71.5	89.7	110.8	118.5
Golden rain	80.0	98.5	114.7	124.8
Marvellous	86.7	108.5	117.2	126.8

## Standard errors of differences

Average:	9.161
Maximum:	9.715
Minimum:	7.683

Average variance of differences: 84.74

Standard error of differences for same level of factor:

	variety	nitrogen
Average:	7.683	9.715
Maximum:	7.683	9.715
Minimum:	7.683	9.715

## Approximate least significant differences (5% level) of REML means

## variety

variety Victory	1	*		
variety Golden rain	2	15.77	*	
variety Marvellous	3	15.77	15.77	*
		1	2	3

## nitrogen

nitrogen 0 cwt	1	*			
nitrogen 0.2 cwt	2	8.934	*		
nitrogen 0.4 cwt	3	8.934	8.934	*	
nitrogen 0.6 cwt	4	8.934	8.934	8.934	*
		1	2	3	4

## variety.nitrogen

Message: caution - t-values using d.d.f. from contributing terms differ by



9.61%; LSD's will be calculated using the maximum value.

variety Victory.nitrogen 0 cwt	1	*		
variety Victory.nitrogen 0.2 cwt	2	17.12	*	
variety Victory.nitrogen 0.4 cwt	3	17.12	17.12	*
variety Victory.nitrogen 0.6 cwt	4	17.12	17.12	17.12
variety Golden rain.nitrogen 0 cwt	5	21.65	21.65	21.65
variety Golden rain.nitrogen 0.2 cwt	6	21.65	21.65	21.65
variety Golden rain.nitrogen 0.4 cwt	7	21.65	21.65	21.65
variety Golden rain.nitrogen 0.6 cwt	8	21.65	21.65	21.65
variety Marvellous.nitrogen 0 cwt	9	21.65	21.65	21.65
variety Marvellous.nitrogen 0.2 cwt	10	21.65	21.65	21.65
variety Marvellous.nitrogen 0.4 cwt	11	21.65	21.65	21.65
variety Marvellous.nitrogen 0.6 cwt	12	21.65	21.65	21.65
		1	2	3
variety Victory.nitrogen 0.6 cwt	4	*		
variety Golden rain.nitrogen 0 cwt	5	21.65	*	
variety Golden rain.nitrogen 0.2 cwt	6	21.65	17.12	*
variety Golden rain.nitrogen 0.4 cwt	7	21.65	17.12	17.12
variety Golden rain.nitrogen 0.6 cwt	8	21.65	17.12	17.12
variety Marvellous.nitrogen 0 cwt	9	21.65	21.65	21.65
variety Marvellous.nitrogen 0.2 cwt	10	21.65	21.65	21.65
variety Marvellous.nitrogen 0.4 cwt	11	21.65	21.65	21.65
variety Marvellous.nitrogen 0.6 cwt	12	21.65	21.65	21.65
		4	5	6
variety Golden rain.nitrogen 0.4 cwt	7	*		
variety Golden rain.nitrogen 0.6 cwt	8	17.12	*	
variety Marvellous.nitrogen 0 cwt	9	21.65	21.65	*
variety Marvellous.nitrogen 0.2 cwt	10	21.65	21.65	17.12
variety Marvellous.nitrogen 0.4 cwt	11	21.65	21.65	17.12
variety Marvellous.nitrogen 0.6 cwt	12	21.65	21.65	17.12
		7	8	9
variety Marvellous.nitrogen 0.2 cwt	10	*		
variety Marvellous.nitrogen 0.4 cwt	11	17.12	*	
variety Marvellous.nitrogen 0.6 cwt	12	17.12	17.12	*
		10	11	12

The least significant differences are calculated using the denominator degrees of freedom (d.d.f.) from the tests for fixed effects. The degrees of freedom are relevant for assessing the fixed term as a whole, and may vary over the contrasts amongst the means of the term. So the LSDs may be only approximate, and should be used with caution. The `variety`  $\times$  `nitrogen` table of means is calculated from the `variety` and `nitrogen` main effects and the `variety.nitrogen` interaction. So, we know that contrasts in the table may have either 10 or 45 d.d.f. Genstat takes the smallest of the available d.d.f. and prints a warning if the resulting LSDs would differ by more than 1%. As a result the LSDs do not match those from the [Analysis of Variance](#) menu, whereas they do for `variety` and `nitrogen` (which have the same d.d.f. for all their contrasts).

The **REML** menus thus produce the same information as that given by the **Analysis of Variance** menu where this is possible in their more general context, but they are not able to match its more specialized output. The advantage of **REML**, however, lies in the much wider range of situations that it covers, as we will show later in this Guide.

There are three menus that you can use to examine the fixed model in more detail, once you have run the initial analysis. They do this by a generalized regression analysis, with a weight matrix based on variances estimated from the original **REML** analysis (with the full fixed model). The models are thus assessed against identical estimates of the random variation (as in an analysis of variance), allowing statistics such as the Akaike information criterion to be used to assess which model may be best. Conversely, if you try to assess the fixed model by changing the fixed model in the original menus, the fixed terms that are not fitted will be included in the random variation. This will then vary from fit to fit, and it may be difficult to reach a clear and consistent conclusion.

The menu for **REML Screening of Fixed Terms**, opened by clicking on the **Screening tests** button in the **Linear Mixed Models Further Output** menu (Figure 1.6), is useful to overcome the problems of ordering that occur when there are non-orthogonal terms. In Figure 1.7, we have chosen to print both Wald tests (which come first in the output, below), as well as F tests.

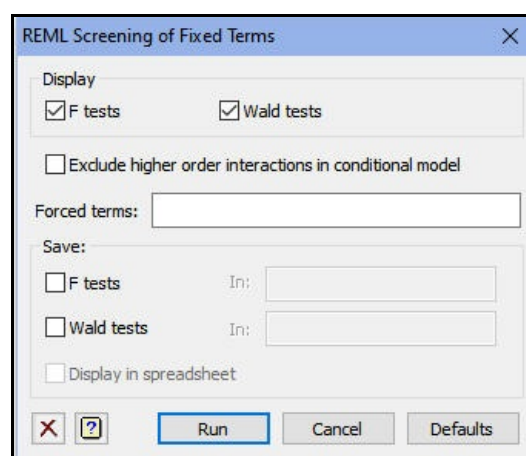


Figure 1.7

## Screening tests for fixed effects

Fixed term	d.f.	Marginal Wald test	pr.	Conditional Wald test	pr.
variety	2	2.97	0.226	2.97	0.226
nitrogen	3	113.06	<0.001	113.06	<0.001
variety.nitrogen	6	1.82	0.936	1.82	0.936

Fixed term	n.d.f.	d.d.f.	Marginal F test	pr.	Conditional F test	pr.
variety	2	10	1.49	0.272	1.49	0.272
nitrogen	3	45	37.69	<0.001	37.69	<0.001
variety.nitrogen	6	45	0.30	0.932	0.30	0.932

In the *marginal tests*, each term is added to the simplest possible model. So in a model with

$$A + B + C + A.B + A.C + B.C + A.B.C$$

$A.B$  would be added to a model containing only the main effects  $A$  and  $B$ . This assesses the effect of the term ignoring as many other terms as possible. It is therefore checking to see if there is *any* evidence that the term has an effect.

In the *conditional tests*, each term is added to the most complex possible model. So,  $A$  would be added to a model containing  $B$ ,  $C$  and  $B.C$ . This checks to see if the term has any effect that *cannot* be explained by any other terms. You can check the box to **Exclude higher order interactions in conditional model** if you would prefer terms containing more factors than the original term not to be included in the model to which the term is added. The **Forced terms** box allows you to specify terms that must always be included in the models.

In this example the terms are orthogonal, and so the marginal and conditional tests are identical. In non-orthogonal analyses they will differ, but the hope is that they lead to the same conclusion. If not, the conclusion is that there is more than one plausible model for the data, but the design is too unbalanced to allow you to choose between them!

Secondly, there is a menu to assess all subsets of the fixed terms, opened by clicking on the **All subsets** button in the **Linear Mixed Models Further Output** menu (Figure 1.6), which may be helpful for larger fixed models. The menu has a box at the top, where you can specify any forced terms that must be included in every subset. The radio buttons in the **Marginal terms** box control whether terms marginal to any other fixed term must be treated as forced and always included. (A term is *marginal* to another term if all its factors or variates also occur in that other term. For example the terms *variety* and *nitrogen* are both marginal to the interaction *variety.nitrogen*.) If you specify that they are to be forced, then all the models will be obtainable by dropping terms that appear in the lower part of the table of tests for fixed effects (i.e. the table obtained by checking the **Wald tests** box in the **Linear Mixed Models Options** or **Linear Mixed Models Further Output** menus). In Figure 1.8 we have specified no forced terms, and selected the Akaike and Schwarz Bayesian information criteria for the output. The results confirm that there is evidence that only nitrogen has an effect on the yields.

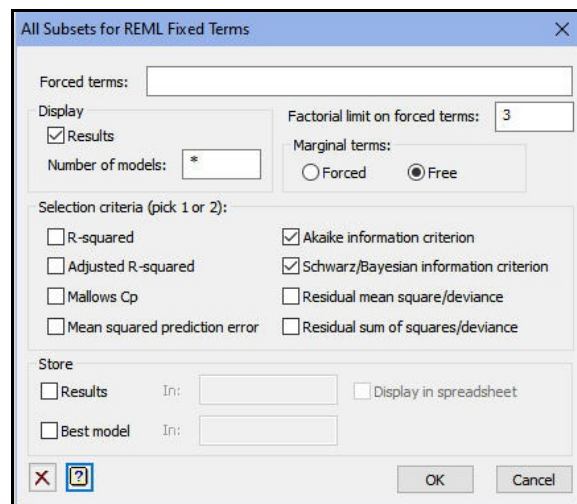


Figure 1.8

## All subsets of REML fixed model

Free terms:

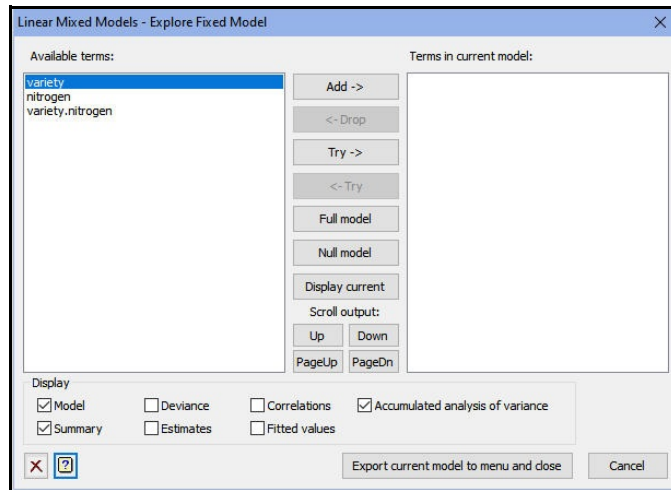
- (1) variety
- (2) nitrogen
- (3) variety.nitrogen

Set	Aic	Sic	d.f.	(1)	(2)	(3)
1	179.8	182.1	1	-	-	-

2	72.8	81.9	4	-	<.001	-
3	180.9	187.7	3	0.272	-	-
4	73.8	87.5	6	0.272	<.001	-
5	84.0	111.3	12	marg	marg	0.932

Set 2 is best according to Aic: nitrogen.  
Set 2 is best according to Sic: nitrogen.

Finally, there is a menu where you can explore the fixed model, by trying, adding and dropping terms in a similar way to the general linear regression menu. (See Section 1.11 of the *Guide to Regression, Nonlinear and Generalized Linear Models*.) You can open this by clicking on the [Explore fixed model](#) button in the [Linear Mixed Models](#) menu. This is greyed out, as Figure 1.6, until you have run the initial analysis.



**Figure 1.9**

The menu has buttons to fit the full model, or to return to the null model. You can also investigate terms individually by clicking on the [Add](#), [Drop](#) or [Try](#) buttons, with the relevant term highlighted in the appropriate window ([Available terms](#) for [Add](#), or [Terms in the current model](#) for [Drop](#), or either window for [Try](#)). The output below shows the effect of clicking on the [Full model](#) button in Figure 1.9.

## Regression analysis of REML fixed model

Response variate: yield  
Weight matrix: REML weights  
Fitted terms: Constant + variety + nitrogen + variety.nitrogen

### Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	11	117.84	10.713	10.71
Residual	60	60.00	1.000	
Total	71	177.84	2.505	
Change	-6	-1.82	0.303	0.30

Percentage variance accounted for 60.1  
Akaike information criterion is estimated to be 318.79.  
Schwarz Bayes information criterion is estimated to be 346.11.

## Accumulated analysis of variance

Change	s.s.	n.d.f.	m.s.	F	d.d.f.	pr.
+ variety	2.97	2	1.485	1.49	10.00	0.272
+ nitrogen	113.06	3	37.686	37.69	45.00	<.001
+ variety.nitrogen	1.82	6	0.303	0.30	45.00	0.932
Residual	60.00	60	1.000			
Total	177.84	71	2.505			

We can highlight `variety.nitrogen` in the **Terms in the current model** window, and click on **Drop** as in Figure 1.10 to remove it from the model.

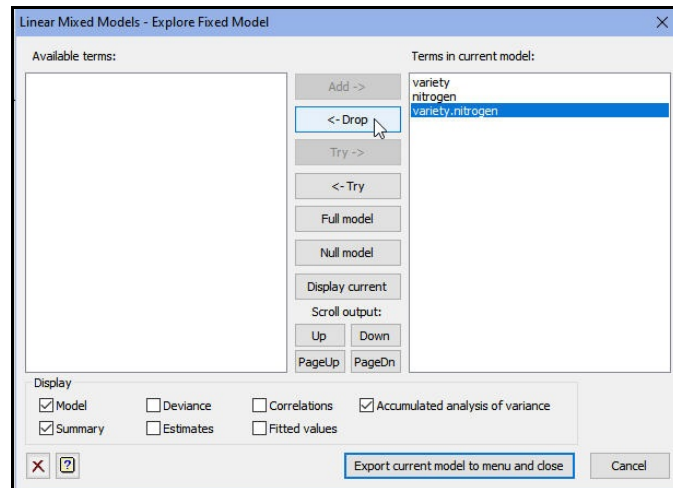


Figure 1.10

## Regression analysis of REML fixed model

Response variate: yield  
 Weight matrix: REML weights  
 Fitted terms: Constant + variety + nitrogen

## Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	5	116.03	23.2055	24.78
Residual	66	61.82	0.9366	
Total	71	177.84	2.5049	
Change	6	1.82	0.3028	0.30

Percentage variance accounted for 62.6  
 Akaike information criterion is estimated to be 308.94.  
 Schwarz Bayes information criterion is estimated to be 322.60.

## Accumulated analysis of variance

Change	s.s.	n.d.f.	m.s.	F	d.d.f.	pr.
+ variety	2.97	2	1.485	1.49	10.00	0.272
+ nitrogen	113.06	3	37.686	37.69	45.00	<.001
+ variety.nitrogen	1.82	6	0.303	0.30	45.00	0.932
Residual	60.00	60	1.000			
- variety.nitrogen	-1.82	-6	0.303	0.30	45.00	0.932
Total	177.84	71	2.505			

---

We can then repeat this for **variety** to obtain a model containing only **nitrogen** (and no non-significant terms).

---

## Regression analysis of REML fixed model

Response variate: yield  
 Weight matrix: REML weights  
 Fitted terms: Constant + nitrogen

## Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	3	113.06	37.6856	39.55
Residual	68	64.79	0.9528	
Total	71	177.84	2.5049	
Change	2	2.97	1.4853	1.59

Percentage variance accounted for 62.0

Akaike information criterion is estimated to be 308.32.

Schwarz Bayes information criterion is estimated to be 317.43.

## Accumulated analysis of variance

Change	s.s.	n.d.f.	m.s.	F	d.d.f.	pr.
+ variety	2.97	2	1.485	1.49	10.00	0.272
+ nitrogen	113.06	3	37.686	37.69	45.00	<.001
+ variety.nitrogen	1.82	6	0.303	0.30	45.00	0.932
Residual	60.00	60	1.000			
- variety.nitrogen	-1.82	-6	0.303	0.30	45.00	0.932
- variety	-2.97	-2	1.485	1.49	10.00	0.272
Total	177.84	71	2.505			

---

You can refit the **REML** analysis with the fixed model redefined to contain only the terms in the current model in this menu, by clicking on the button to [Export current model to menu and close](#). An alternative strategy would be to keep the original fixed model, and form predictions using only the terms in this current model (with the [REML Predictions](#) menu described in Section 1.6).

## 1.2 Commands for REML analysis

REML analysis in Genstat is performed by the `REML` directive. Before using `REML` we first need to define the model that is to be fitted in the analysis. For straightforward linear mixed models, the only directive that needs to be specified is `VCOMPONENTS`. We give the full syntax below, and then highlight the main options and parameters. The full details are in the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 5.2.1.

---

### **VCOMPONENTS directive**

Defines the variance-components model for `REML`.

#### **Options**

<code>FIXED</code> = <i>formula</i>	Fixed model terms; default *
<code>ABSORB</code> = <i>factor</i>	Defines the absorbing factor (appropriate only when <code>REML</code> option <code>METHOD=Fisher</code> ); default * i.e. none
<code>CONSTANT</code> = <i>string token</i>	How to treat the constant term ( <i>estimate</i> , <i>omit</i> ); default <i>esti</i>
<code>FACTORIAL</code> = <i>scalar</i>	Limit on the number of factors or covariates in each fixed term; default 3
<code>CADJUST</code> = <i>string token</i>	What adjustment to make to covariates before analysis ( <i>mean</i> , <i>none</i> ); default <i>mean</i>
<code>RELATIONSHIP</code> = <i>matrix</i>	Defines relationships constraining the values of the components; default *
<code>SPLINE</code> = <i>formula</i>	Defines random cubic spline terms to be generated: each term must contain only one variate, if there is more than one factor in a term, separate splines are calculated for each combination of levels of the factors
<code>EXPERIMENTS</code> = <i>factor</i>	Factor defining the different experiments in a multi-experiment (meta-) analysis

#### **Parameters**

<code>RANDOM</code> = <i>formula</i>	Random model terms
<code>INITIAL</code> = <i>scalars</i>	Initial values for each component and the residual variance
<code>CONSTRAINTS</code> = <i>string tokens</i>	How to constrain each variance component and the residual variance ( <i>none</i> , <i>positive</i> , <i>fixrelative</i> , <i>fixabsolute</i> ); default <i>none</i>

---

The `FIXED` option specifies a model formula defining the fixed model terms to be fitted, and the `RANDOM` parameter specifies another model formula defining the random terms. There are two other parameters. `INITIAL` provides initial values for the estimation of each variance component. These are supplied as the ratio of the component to the residual variance, but the default value of one is usually satisfactory. The `CONSTRAINT` parameter can be used to indicate whether each variance component is to be constrained in any way.



The default setting, `none`, leaves them unconstrained. The `positive` setting forces a variance component to be kept positive, the `fixrelative` fixes the relative value of the component to be equal to that specified by the `INITIAL` parameter, and the `fixabsolute` setting fixes it to the absolute value specified by `INITIAL`. The `FACTORIAL` option sets a limit on the number of factors and variates allowed in each fixed term (default 3); any term containing more than that number is deleted from the model.

Usually, only `FIXED` and `RANDOM` need to be set. For example, the statement below defines the models for the split-plot example in Section 1.2.

```
VCOMPONENTS [FIXED=variety*nitrogen] \
RANDOM=blocks/wplots/subplots
```

Once the models have been defined, the `REML` directive can be used to perform the analysis.

---

## REML directive

Fits a variance-components model by residual (or restricted) maximum likelihood.

### Options

<code>PRINT</code> = <i>string tokens</i>	What output to present ( <code>model</code> , <code>components</code> , <code>effects</code> , <code>means</code> , <code>stratumvariances</code> , <code>monitoring</code> , <code>vcovariance</code> , <code>deviance</code> , <code>Waldtests</code> , <code>missingvalues</code> , <code>covariancemodels</code> ); default <code>mode</code> , <code>comp</code> , <code>Wald</code> , <code>cova</code>
<code>PTERMS</code> = <i>formula</i>	Terms (fixed or random) for which effects or means are to be printed; default <code>*</code> implies all the fixed terms
<code>PSE</code> = <i>string token</i>	Standard errors to be printed with tables of effects and means ( <code>differences</code> , <code>estimates</code> , <code>alldifferences</code> , <code>allestimates</code> , <code>none</code> ); default <code>diff</code>
<code>WEIGHTS</code> = <i>variate</i>	Weights for the analysis; default <code>*</code> implies all weights 1
<code>MVINCLUDE</code> = <i>string tokens</i>	Whether to include units with missing values in the explanatory factors and variates and/or the y-variates ( <code>explanatory</code> , <code>yvariate</code> ); default <code>*</code> i.e. omit units with missing values in either explanatory factors or variates or y-variates
<code>SUBMODEL</code> = <i>formula</i>	Defines a submodel of the fixed model to be assessed against the full model (for <code>METHOD=Fisher</code> only)
<code>RECYCLE</code> = <i>string token</i>	Whether to reuse the results from the estimation when printing or assessing a submodel ( <code>yes</code> , <code>no</code> ); default <code>no</code>
<code>RMETHOD</code> = <i>string token</i>	Which random terms to use when calculating



	RESIDUALS (final, all, notspline); default final
METHOD = <i>string token</i>	Indicates whether to use the standard Fisher-scoring algorithm or the new AI algorithm with sparse matrix methods (Fisher, AI); default AI
MAXCYCLE = <i>scalar</i>	Limit on the number of iterations; default 30
TOLERANCES = <i>variate</i>	Tolerances for matrix inversion; default * i.e. appropriate default values
PARAMETERIZATION = <i>string token</i>	Parameterization to use for the variance component estimation (gammas, sigmas)
CFORMAT = <i>string token</i>	Whether printed output for covariance models gives the variance matrices or the parameters (variancematrices, parameters); default vari
FMETHOD = <i>string token</i>	Controls whether and how to calculate F-statistics for fixed terms (automatic, none, algebraic, numerical); default auto
WORKSPACE = <i>scalar</i>	Number of blocks of internal memory to be allocated for use by the estimation algorithm when METHOD=AI

### Parameters

Y = <i>variates</i>	Variates to be analysed
RESIDUALS = <i>variates</i>	Residuals from each analysis
FITTEDVALUES = <i>variates</i>	Fitted values from each analysis
SAVE = <i>pointers</i>	Saves the details of each analysis for use in subsequent VDISPLAY and VKEEP directives

The first parameter of REML specifies the y-variate to be analysed. The PRINT option is set to a list of strings to select the output to be printed. These are similar to the check boxes of the [Linear Mixed Models Further Output](#) menu. The most commonly used settings are:

model	description of model fitted,
components	estimates of variance components and estimated parameters of covariance models,
effects	estimates of parameters in the fixed and random models,
means	predicted means for factor combinations,
vcovariance	variance-covariance matrix of the estimated components,
deviance	deviance of the fitted model,
waldtests	Wald tests for all fixed terms in model,
missingvalue	estimates of missing values,
covariancemodels	estimated covariance models.

The default setting of PRINT=model, components, Wald, cov gives a description of the model and covariance models that have been fitted, together with estimates of the variance components and the Wald (or F) tests. By default if tables of means and effects

are requested, tables for all terms in the fixed model are given together with a summary of the standard error of differences between effects/means.

The `FMETHOD` option controls whether and how the F statistics are calculated, with settings `automatic`, `none`, `algebraic` and `numerical`.

The `PTERMS` and `PSE` options can be used to change the terms or obtain different types of standard error. For example,

```
REML [PRINT=means; PTERMS=nitrogen.variety; \
      PSE=alallestimates]
```

will produce a nitrogen by variety table of predicted means with a standard error for each cell.

Further output is produced by the `VDISPLAY` directive, which has options `PRINT`, `PTERMS` and `PSE` like those of `REML`.

---

### **VDISPLAY directive**

Displays further output from a `REML` analysis.

#### **Options**

<code>PRINT</code> = <i>string tokens</i>	What output to present ( <code>model</code> , <code>components</code> , <code>effects</code> , <code>means</code> , <code>stratumvariances</code> , <code>monitoring</code> , <code>vcovariance</code> , <code>deviance</code> , <code>Waldtests</code> , <code>missingvalues</code> , <code>covariancemodels</code> ); default <code>mode</code> , <code>comp</code> , <code>Wald</code> , <code>cova</code>
<code>CHANNEL</code> = <i>identifier</i>	Channel number of file, or identifier of a text to store output; default current output file
<code>PTERMS</code> = <i>formula</i>	Terms (fixed or random) for which effects or means are to be printed; default <code>*</code> implies all the fixed terms
<code>PSE</code> = <i>string token</i>	Standard errors to be printed with tables of effects and means ( <code>differences</code> , <code>estimates</code> , <code>alldifferences</code> , <code>alallestimates</code> , <code>none</code> ); default <code>diff</code>
<code>CFORMAT</code> = <i>string token</i>	Whether printed output for covariance models gives the variance matrices or the parameters ( <code>variancematrices</code> , <code>parameters</code> ); default <code>vari</code>
<code>FMETHOD</code> = <i>string token</i>	Controls whether and how to calculate F-statistics for fixed terms ( <code>automatic</code> , <code>none</code> , <code>algebraic</code> , <code>numerical</code> ); default <code>auto</code>

#### **Parameter**

<i>pointers</i>	Save structure containing the details of each analysis; default is to take the save structure from the latest <code>REML</code> analysis
-----------------	--

---

The approximate least significant differences are calculated by the `VLSD` procedure.

**VLSD procedure**

Prints approximate least significant differences for REML means (R.W. Payne).

**Options**

<code>PRINT</code> = <i>string tokens</i>	Controls printed output ( <code>means</code> , <code>sed</code> , <code>lsd</code> , <code>df</code> ); default <code>lsd</code>
<code>FACTORIAL</code> = <i>scalar</i>	Limit on the number of factors in each term; default 3
<code>LSDLEVEL</code> = <i>scalar</i>	Significance level (%) to use in the calculation of least significant differences; default 5
<code>DFMETHOD</code> = <i>string token</i>	Specifies which degrees of freedom to use for the t-statistics ( <code>fddf</code> , <code>given</code> , <code>tryfddf</code> ); default <code>fddf</code>
<code>DFGIVEN</code> = <i>scalar</i>	Specifies the number of degrees of freedom to use for the t-statistics when <code>DFMETHOD=given</code> , or if d.d.f. are unavailable when <code>DFMETHOD=tryfddf</code>
<code>FMETHOD</code> = <i>string token</i>	Controls how to calculate denominator degrees of freedom for the F-statistics, if these are not already available in the REML save structure ( <code>automatic</code> , <code>algebraic</code> , <code>numerical</code> ); default <code>auto</code>
<code>SAVE</code> = <i>REML save structure</i>	Save structure to provide the table of means; default uses the save structure from the most recent REML

**Parameters**

<code>TERMS</code> = <i>formula</i>	Treatment terms whose means are to be compared; default <code>*</code> takes the REML fixed model
<code>MEANS</code> = <i>pointer or table</i>	Saves the means for each term
<code>SED</code> = <i>pointer or symmetric matrix</i>	Saves standard errors of differences between means
<code>LSD</code> = <i>pointer or symmetric matrix</i>	Saves approximate least significant differences matrix for the means
<code>DF</code> = <i>pointer or scalar</i>	Saves the degrees of freedom used to calculate the t critical values for the LSDs
<code>DDF</code> = <i>pointer or scalar</i>	Saves the denominator degrees of freedom in the F test for the term
<code>DFRANGE</code> = <i>pointer or scalar</i>	Saves the range of denominator degrees of freedom in the F tests for the term and any terms that are marginal to the term (available only when denominator degrees of freedom of F-statistics are being used)

---

VLSD calculates least significant differences (LSDs) for predicted means of fixed terms

in a `REML` analysis. These are calculated by multiplying standard errors for differences by the t-statistic that would be used to assess whether those differences are non-zero.

The `TERMS` parameter specifies a model formula to define the fixed terms whose predicted means are to be compared. The means are usually taken from the most recent analysis performed by `REML`, but you can set the `SAVE` option to a save structure from another `REML` if you want to examine means from an earlier analysis. As in `VCOMPONENTS`, the `FACTORIAL` option sets a limit on the number of factors in each term (default 3).

The `DFMETHOD` option specifies how to obtain the degrees of freedom for the t-statistics. The default is to use the numbers of denominator degrees of freedom printed by `REML` in the `d.d.f.` column in the table of tests for fixed tests (produced by setting option `PRINT=wald`). The degrees of freedom are relevant for assessing the fixed term as a whole, and may vary over the contrasts amongst the means of the term. So the LSDs should be used with caution. (If you are interested in a specific comparison, you should set up a 2-level factor to fit this explicitly in the analysis.) The `FMETHOD` option controls how the denominator degrees of freedom should be calculated, if they are not already available in the `REML` save structure (e.g. because they were printed in the original analysis). The settings are the same as in the `REML` and `VKEEP` directives, except that there is no `none` setting. (You would set this option only if you really do want to calculate them.)

In some of the more complicated analyses, `REML` may be unable to calculate the denominator degrees of freedom. You might then want to supply the number of degrees of freedom yourself, using the `DFGIVEN` option, rather than having no least significant differences at all. For example, you could use the number of denominator degrees of freedom from the analysis of an earlier similar design. However, the results will only be as good as the degrees of freedom that you have supplied, and thus should be used with caution! You can set option `DFMETHOD=tryfddf` to use the denominator degrees of freedom, if these can be calculated, or those specified by `DFGIVEN` otherwise. The setting `DFMETHOD=given` always uses the degrees of freedom specified by `DFGIVEN`.

Printed output is controlled by the `PRINT` option, with settings:

<code>means</code>	prints the means;
<code>sed</code>	prints standard errors for differences between the means;
<code>lsd</code>	prints least significant differences for the means;
<code>df</code>	prints the degrees of freedom used to calculate the t critical value required for the LSD, together with the denominator degrees of freedom in the F test for the term if these are not the same.

The significance level to use in the calculation of the least significant differences can be changed from the default of 5% using the `LSDLEVEL` option.

The `MEANS` parameter can save the means. If the `TERMS` parameter specifies a single term, `MEANS` must be undeclared or set to a table. If `TERMS` specifies several terms, you must supply a pointer which will then be set up to contain as many tables as there are terms. Similarly the `SED` parameter can save the standard errors of differences, the `LSD` parameter can save the approximate least significant differences, the `DF` parameter can save the degrees of freedom used to calculate the t-statistics, and the `DDF` parameter can save the denominator degrees of freedom in the F tests.

When a term involves several factors, its means may be formed from the effects of several terms. For example, the means for the term `A.B` will involve the effects for the terms `A` and `B` (if these are in the model), as well as those for the term `A.B`. Different contrasts between the means will then have different denominator degrees of freedom. For caution, if `VLSD` is using the number of denominator degrees of freedom, it uses the smallest number over the terms that are involved in calculating each table of the means. (This corresponds to the largest t-statistic.) If the difference in the t-statistics calculated from smallest and largest numbers of degrees of freedom differ by more than 1%, `VLSD` prints a warning message. If the denominator degrees of freedom are being used, their range for each term can be saved by the `DFRANGE` parameter.

The screening tests are produced by the `VSCREEN` procedure. Details are in the *Genstat Reference Manual, Part 3 Procedures*.

### 1.3 Practical

In an experiment to study the effect of two meat-tenderizing chemicals, the two (back) legs were taken from four carcasses of beef and one leg was treated with chemical 1 and the other with chemical 2. Three sections were then cut from each leg and allocated (at random) to three cooking temperatures, all 24 sections ( $4 \text{ carcasses} \times 2 \text{ legs} \times 3 \text{ sections}$ ) being cooked in separate ovens. The table below shows the force required to break a strip of meat taken from each of the cooked sections (the data are also in the file `Meat.gsh`). Analyse the experiment using both `ANOVA` and `REML`, and compare the results.

Leg		1			2		
Carcass	Section	Chemical	Temp	Force	Chemical	Temp	Force
1	1	1	2	5.5	2	3	6.3
	2	1	3	6.5	2	1	3.5
	3	1	1	4.3	2	2	4.8
2	1	2	1	3.2	1	3	6.2
	2	2	3	6.0	1	2	5.0
	3	2	2	4.7	1	1	4.0
3	1	2	1	2.6	1	2	4.6
	2	2	2	4.3	1	1	3.8
	3	2	3	5.6	1	3	5.8
4	1	1	3	5.7	2	2	4.1
	2	1	1	3.7	2	3	5.9
	3	1	2	4.9	2	1	2.9

Run screening tests of the fixed terms. Notice that the marginal and conditional tests give the same values. (This is because the design is orthogonal.)

## 1.4 Means plots

It is often interesting to plot the means from a mixed-model analysis. You can open the **Means Plots** menu, by clicking on the **Means plot** button in the **Linear Mixed Model Further Output** menu (Figure 1.6).

The menu (Figure 1.8) allows you to choose a factor to plot along the x-axis. The **Groups** box specifies factors whose levels are to be included in a single window of the graph; Genstat then plots a set of points or a line (depending on your choice in the **Method** box) for each level of the groups factor. The **Trellis groups** box can specify factors to define a trellis plot, and the **Page groups** box can specify factors whose combinations of levels are to be plotted on different pages.

In Figure 1.9, we have plotted the nitrogen-by-variety means, with nitrogen along the x-axis.

The plotting is done using the **VGRAPH** procedure.

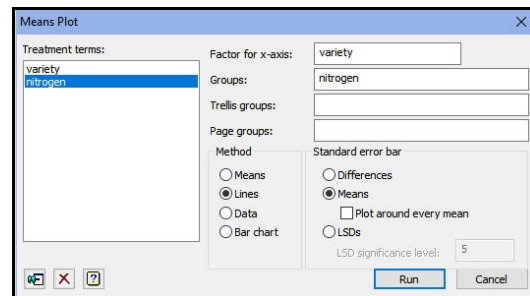


Figure 1.11

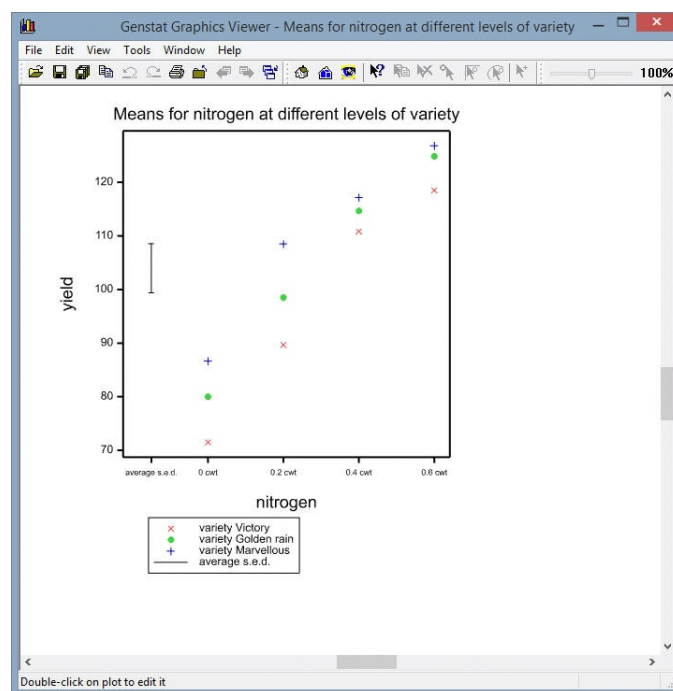


Figure 1.12

**VGRAPH procedure**

Plots tables of means from REML (R.W. Payne).

**Options**

GRAPHICS = <i>string token</i>	Type of graph ( <i>highresolution</i> , <i>lineprinter</i> ); default <i>high</i>
METHOD = <i>string token</i>	What to plot ( <i>points</i> , <i>means</i> , <i>linesandpoints</i> , <i>onlylines</i> , <i>data</i> , <i>barchart</i> , <i>splines</i> ); default <i>poin</i> when XFACTOR is a factor, and <i>only</i> when it is a variate
XFREPRESENTATION = <i>string token</i>	How to label the x-axis ( <i>levels</i> , <i>labels</i> ); default <i>labe</i> uses the XFACTOR labels, if available
PSE = <i>string token</i>	What to plot to represent variation when points are plotted at the means ( <i>differences</i> , <i>lsd</i> , <i>means</i> , <i>allmeans</i> ); default <i>diff</i>
LSDLEVEL = <i>scalar</i>	Significance level (%) to use for approximate least significant differences; default 5
DFSPLINE = <i>scalar</i>	Number of degrees of freedom to use when METHOD= <i>splines</i>
YTRANSFORM = <i>string tokens</i>	Transformed scale for additional axis marks and labels to be plotted on the right-hand side of the y-axis ( <i>identity</i> , <i>log</i> , <i>log10</i> , <i>logit</i> , <i>probit</i> , <i>cloglog</i> , <i>square</i> , <i>exp</i> , <i>exp10</i> , <i>ilogit</i> , <i>iprobit</i> , <i>icloglog</i> , <i>root</i> ); default <i>iden</i> i.e. none
PENYTRANSFORM = <i>scalar</i>	Pen to use to plot the transformed axis marks and labels; default <i>*</i> selects a pen, and defines its properties, automatically
SAVE = <i>REML save structure</i>	Save structure to provide the table of means; default uses the save structure from the most recent REML

**Parameters**

XFACTOR = <i>factors or variates</i>	Provides the x-values for each plot; by default this is chosen automatically
GROUPS = <i>factors or pointers</i>	Factor or factors identifying groups in each plot; by default chosen automatically
TRELLISGROUPS = <i>factors or pointers</i>	Factor or factors specifying the different plots of a trellis plot of a multi-way table
PAGEGROUPS = <i>factors or pointers</i>	Factor or factors specifying plots to be displayed on different pages
NEWXLEVELS = <i>variates</i>	Values to be used for XFACTOR; default uses the existing levels if XFACTOR is a factor, and the



<code>TITLE = texts</code>	minimum and maximum values if it is a variate Title for the graph; default is to define a title automatically if <code>GROUPS</code> is set, or to have none if it is unset
<code>YTITLE = texts</code>	Title for the y-axis; default ' '
<code>XTITLE = texts</code>	Title for the x-axis; default is to use the identifier of the <code>XFACTOR</code>
<code>PENS = variates</code>	Defines the pen to use to plot the points and/or line for each group defined by the <code>GROUPS</code> factors

---

In its simplest form, the behaviour of `VGRAPH` depends on the model. If the fixed model contains only main effects, it plots the means for the first factor in the fixed model. Otherwise, it looks for the first fixed term involving two factors; it then plots the means with one of these factors as the x-axis, and the second as a grouping factor with levels identified by different plotting colours and symbols. The `GRAPHICS` option controls whether a high-resolution or a line-printer graph is plotted; by default `GRAPHICS=high`.

By default, the means are from the most recent `REML`. However, you can plot means from an earlier analysis, by using the `SAVE` option of `VGRAPH` to specify its save structure (saved using the `SAVE` parameter of the `REML` command that performed the analysis). `VGRAPH` uses the `VPREDICT` directive with default option settings to obtain the means (see Section 1.6). This should give the same means as those printed by `REML` or `VDISPLAY`. If you want to use `VPREDICT` with other option settings, you can plot these by saving the predictions, and then using the `DTABLE` procedure.

The `METHOD` option controls how the predicted means are plotted in high-resolution graphics, with settings:

<code>points</code>	to plot a point at each mean;
<code>means</code>	synonym of <code>points</code> ;
<code>linesandpoints</code>	to plot points and join them by lines;
<code>onlylines</code>	to draw lines between the means;
<code>data</code>	to draw lines between the means, and then also plot the original data values;
<code>barchart</code>	to plot the means as a barchart;
<code>splines</code>	to plot points at the means together with a smooth spline to show the trend over each group of means; the <code>DFSPLINE</code> specifies the degrees of freedom for the splines; if this is not set, 2 d.f. are used when there are up to 10 points, 3 if there are 11 to 20, and 4 for 21 or more.

The default is to plot `points` when `XFACTOR` is a factor, and `onlylines` when it is a variate. Only `points` are available in line-printer graphics.

The `PSE` option specifies the type of error bar to be plotted, when points are plotted for the means, with settings:

<code>differences</code>	average standard error of difference;
<code>lsd</code>	average approximate least significant difference (calculated using the <code>VLSD</code> procedure);
<code>means</code>	average effective standard error for the means;



`allmeans` plots plus and minus the effective standard error around every mean.

The `LSDLEVEL` option sets the significance level (%) to use for the approximate least significant differences (default 5). The `allmeans` setting is often unsuitable for plots other than barcharts when there are `GROUPS`, as the plus/minus e.s.e. bars may overlap each other.

You can define the table of means to plot explicitly, by specifying its classifying factors using the `XFACTOR`, `GROUPS`, `TRELLISGROUPS` and `PAGEGROUPS` parameters. The `XFACTOR` parameter can define a factor against whose levels the means are plotted. It can also specify a variate, and `VPREDICT` then sets up a factor automatically, to classify the table, with levels at the values specified by the `NEWXLEVELS` parameter. With a multi-way table, there will be a plot of means against the `XFACTOR` levels for every combination of levels of the factors specified by the `GROUPS`, `TRELLISGROUPS` and `PAGEGROUPS` parameters. The `GROUPS` parameter specifies factors whose levels are to be included in a single window of the graph.

For example, Figure 1.9 was plotted by the statement

```
VGRAPH [METHOD=means; PSE=differences] \
nitrogen; GROUPS=variety
```

The means are plotted in a single window with factor `nitrogen` on the x-axis, and different symbols used for the means with each level of the factor `variety`. You can set `GROUPS` to a pointer to specify several factors to define groups. For example

```
POINTER [VALUES=variety,blocks] Groupfactors
VGRAPH [METHOD=line] nitrogen; GROUPS=Groupfactors
```

Here we have also set option `METHOD=line`. So this plots a line for every combination of the levels of factors `variety` and `blocks`.

Similarly, the `TRELLISGROUPS` option can specify one or more factors to define a trellis plot. For example,

```
VGRAPH [METHOD=line] nitrogen; GROUPS=variety;
TRELLISGROUPS=blocks
```

will produce a plot for each block, in a trellis arrangement; each plot will again have factor `nitrogen` on the x-axis, and a line for every variety. Likewise, the `PAGEGROUPS` parameter can specify factors whose combinations of levels are to be plotted on different pages. So

```
VGRAPH [METHOD=line] nitrogen; GROUPS=variety; PAGEGROUPS=blocks
```

will again produce a plot for block, but now on separate pages. Multi-way tables can be plotted even if the corresponding model term was not in the `REML` analysis. For example you can plot a two-way table even if the analysis contained only the main effects of the two factors; however, the lines will then all be parallel and no LSDs can be included.

The `NEWXLEVELS` parameter enables different levels to be supplied for an `XFACTOR` factor, if its existing levels are unsuitable. If the factor has labels, these are used to label the x-axis unless you set option `XFREPRESENTATION=levels`. When `XFACTOR` is a variate, `NEWXLEVELS` can specify the values where the predictions are to be made. By default, they are made at its minimum and maximum values.

Note that the values predicted by `VPREDICT`, for an `XFACTOR` variate, will not include any spline effects, nor can it take account of any relationships between different variates

in the model. (For example, the model may include a variate and its square.) To take account of relationships like these, you should use `VPREDICT` directly, specifying the linked variables with the `PARALLEL` parameter. Save the table of predictions, and then plot it using `DTABLE`.

The `TITLE`, `YTITLE` and `XTITLE` parameters can supply titles for the graph, the y-axis and the x-axis, respectively. The symbols, colours and line styles that are used in a high-resolution plot are usually set up by `VGRAPH` automatically. If you want to control these yourself, you should use the `PEN` directive to define a pen with your preferred symbol, colour and line style, for each of the groups defined by combinations of the `GROUPS` factors. The pen numbers should then be supplied to `VGRAPH`, in a variate with a value for each group, using the `PENS` parameter.

The `YTRANSFORM` option allows you to include additional axis markings, transformed onto another scale, on the right-hand side of the y-axis. This is useful if your y-variate has been transformed. Suppose, for example, suppose you have analysed a variate of percentages that have been transformed to logits. You might then set `YTRANSFORM=ilogit` (the inverse-logit transformation) to include markings in percentages alongside the logits. You can control the colours of the transformed marks and labels, by defining a pen with the required properties, and specifying it with the `PENYTRANSFORM` option. Otherwise, the default is to plot them in blue.

## 1.5 Practical

Plot the means from the meat-tenderizing example in Practical 1.3.

## 1.6 Predictions

The predicted means that are plotted by the `Means Plots` menu, or printed by the `Options` or `Further Output` menus, are formed by inserting the estimated effects of the relevant terms into the linear model. Terms that do not contribute to the table of means are ignored.

Remember that the model for the split-plot design was

$$y_{ijk} = \mu + v_r + a_s + va_{rs} + b_i + w_{ij} + \varepsilon_{ijk}$$

where

$y_{ijk}$  is the yield from block  $i$ , whole-plot  $j$ , subplot  $k$ ;

the fixed terms in the model are

$\mu$  the overall constant (grand mean),

$v_r$  the main effect of variety  $r$  (where  $r$  is the variety assigned to unit  $ijk$ ),

$a_s$  the main effect of nitrogen application at level  $s$  (where  $s$  is the nitrogen level applied on unit  $ijk$ ), and

$va_{rs}$  their interaction.

and the random model terms are

$b_i$  the effect of block  $i$ ,

$w_{ij}$  the effect of whole-plot  $j$  within block  $i$ , and

$\varepsilon_{ijk}$  the random error for unit  $ijk$  (which here is the same as the subplot effect, since the subplots are the smallest units of the experiment).

So, to form the nitrogen table of means we would need the grand mean  $\mu$  and the nitrogen effects  $a_s$ . The other effects are omitted. So, equivalently, we can say that we have taken

the random effects at their population means of zero.

Unlike ANOVA, REML allows you to produce predicted values that include estimated values of effects in random terms. (So you are including the estimated effects for the specific sample, in this experiment, from the underlying random population represented by the random term.) These random estimates are known as *best linear unbiased predictors* (or BLUPs). An interesting feature is that the BLUP estimates for a random term are usually smaller than the estimates that would have been obtained if the term had been estimated as a fixed term. For this reason, the BLUP estimates are often called “shrunk” parameter estimates. Further details, and the underlying mathematics, are given in the *Guide to the Genstat Command Language Part 2 Statistics*, Section 5.3.3.

The REML Predictions menu (Figure 1.10) provides more flexibility. To open the menu, you click on the Predict button on the Linear Mixed Models menu (Figure 1.4). This uses the VPREDICT directive, which offers similar facilities to the PREDICT directive in regression analysis. So the predictions are formed by two steps.

- 1 a table of fitted values is calculated. By default these are formed using all the fixed terms in the model, and any random terms that are involved in the prediction table. However, you can specify the terms explicitly by checking the Specify model terms box, and clicking on the Model terms button.
- 2 the fitted values are averaged over the factors that are not in the prediction table. You can control how the averaging is done by clicking on the Averaging options button.

In Figure 1.10 we are constructing predictions for nitrogen with the default settings, except that we are asking to display (all) SEDs instead of the default of Average SEDs. VPREDICT thus forms a nitrogen-by-variety table of fitted values using only the fixed effects  $\mu$ ,  $a_s$  and  $v_r$ . It then averages over variety (using equal weighting) to get the predictions. As the varieties had equal replication in the design, this gives the same results as the ordinary predicted means, shown in Section 1.2.

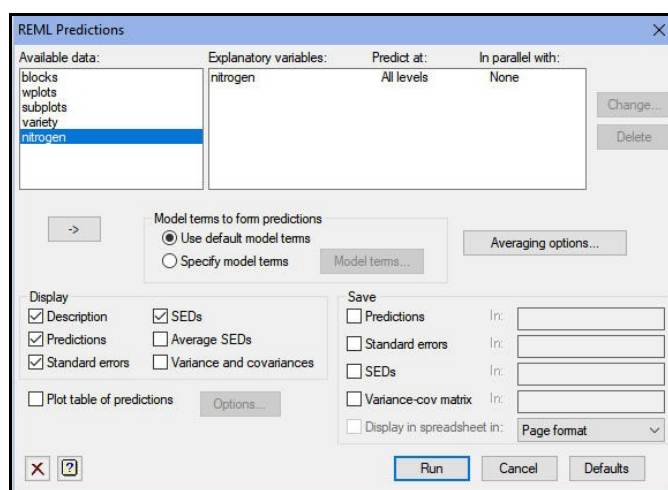


Figure 1.10

## Predictions from REML analysis

Model terms included for prediction: Constant + variety + nitrogen + variety.nitrogen

Model terms excluded for prediction: blocks + blocks.wplots

Status of model variables in prediction:

Variable	Type	Status
nitrogen	factor	Classifies predictions
variety	factor	Averaged over - equal weights
Constant	factor	Included in prediction
blocks	factor	Ignored
wplots	factor	Ignored

Response variate: yield

Predictions

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	79.4	98.9	114.2	123.4

Standard errors

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	7.175	7.175	7.175	7.175

Standard error of differences

nitrogen 0 cwt	1		*		
nitrogen 0.2 cwt	2	4.436		*	
nitrogen 0.4 cwt	3	4.436	4.436		*
nitrogen 0.6 cwt	4	4.436	4.436	4.436	*
		1	2	3	4

These predictions are called *marginal predictions*), as the random effects have been set at their population means of zero.

Another possibility would be to make predictions *conditional* on the block and whole-plot random terms. We would then average over the block and whole-plot (BLUP) effects actually estimated in the analysis. Essentially this predicts the average yields that would be obtained for each level of nitrogen if we ran the experiment again with the same treatment combinations on the

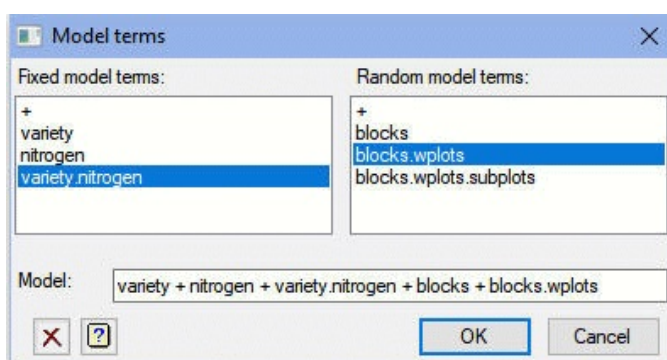


Figure 1.11

plots, and the same block and whole-plot samples from their respective populations. We now need to define the model terms used to make the prediction explicitly, and include `blocks` and `blocks.wplots` as well as the fixed terms. So, we check the [Specify model terms](#) box and click on the [Model terms](#) button in the [REML Predictions](#) menu (Figure 1.10). In the [Model terms](#) menu (Figure 1.11), we then define the model to be

```
variety +nitrogen +variety.nitrogen +blocks +blocks.wplots
```

The output is shown below.

## Predictions from REML analysis

Model terms included for prediction: Constant + variety + nitrogen + variety.nitrogen + blocks + blocks.wplots

Status of model variables in prediction:

Variable	Type	Status
nitrogen	factor	Classifies predictions
variety	factor	Averaged over - equal weights
Constant	factor	Included in prediction
blocks	factor	Averaged over - equal weights
wplots	factor	Averaged over - equal weights

Response variate: yield

Predictions

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	79.4	98.9	114.2	123.4

Standard errors

nitrogen	0 cwt	0.2 cwt	0.4 cwt	0.6 cwt
	3.137	3.137	3.137	3.137

Standard error of differences

nitrogen 0 cwt	1	*				
nitrogen 0.2 cwt	2	4.436	*			
nitrogen 0.4 cwt	3	4.436	4.436	*		
nitrogen 0.6 cwt	4	4.436	4.436	4.436	*	
		1	2	3	4	

In this case, because all the design is orthogonal with factor levels equally replicated, the actual predictions remain the same. However, the standard errors have decreased, as adding information about random effects decreases uncertainty (you would use these to compare a mean with zero). The standard errors of differences are unchanged as we have simply added a constant to every prediction, and this cancels when you take differences. (The constant is the average of the block and whole-plot effects,  $b_i$  and  $w_{ij}$ , which is zero here – but it could be non-zero in a more complicated design.)

Next we show how to use unequal weights in step 2. We click on the **Averaging options** button in the **REML Predictions** menu (Figure 1.10) to open the **Averaging Terms** menu. You might want to check the **Take averages of combinations present** box (as in Figure 1.12) if you

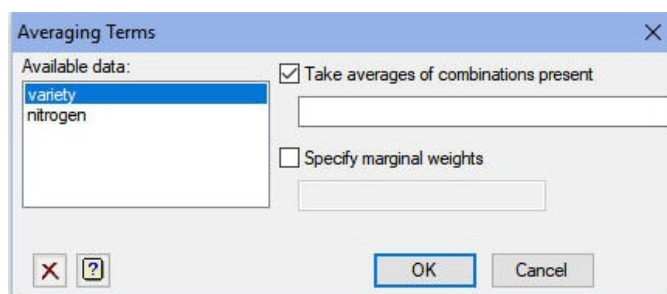


Figure 1.12

have any factors that have some levels with zero replication in the experiment. You can list these in the box below the check box to tell **VPREDICT** to omit these levels during the averaging.

Alternatively, if you check the **Specify marginal weights** box, you can specify a one-way table, classified by any of the averaged factors, to define the weights explicitly. As an example, we will form predictions for varieties assuming that there will always be some nitrogen fertilizer.

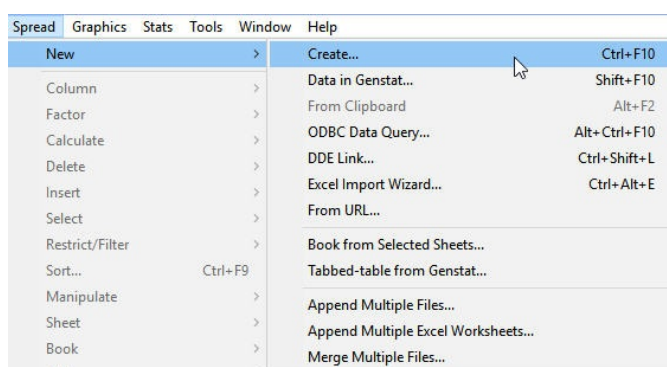


Figure 1.13

We first create the table using the spreadsheet menus. We select the **Create** sub-option of the **New** option of the **Spread** menu on the menu bar, as shown in Figure 1.13.

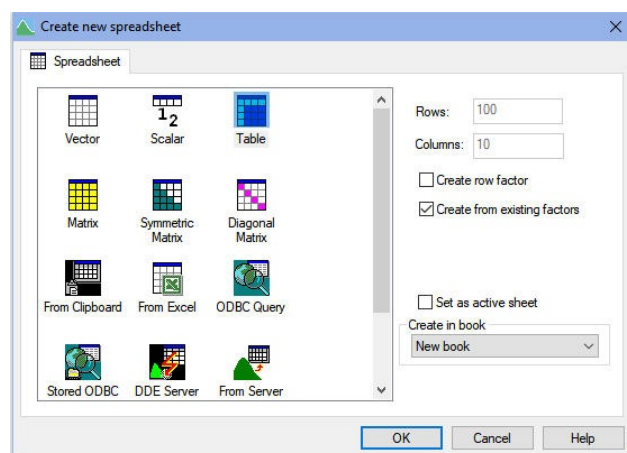


Figure 1.14

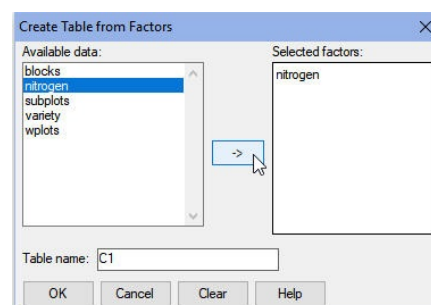


Figure 1.15

This opens the **Size of New Spreadsheet** menu, where we select **Table** as the type of spreadsheet, and check the **Create from existing factors** box (Figure 1.14). We then get the **Create Table from Factors** menu, where we select **nitrogen** as the factor (Figure 1.15).

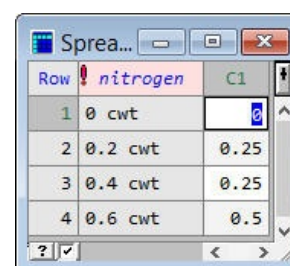


Figure 1.16



The resulting spreadsheet is shown in Figure 1.16. We have left its name as the default, **C1**, chosen by the **Spread** menus, and filled in weights to exclude the zero level of nitrogen, to use 0.25 for 0.2 cwt and 0.4 cwt, and 0.5 for 0.6 cwt. Notice that we have arranged for the weights to sum to one. **VPREDICT** will not do this automatically, but will use the actual weights that you supply.

We can now select **variety** as the explanatory variable in the **REML Predictions** menu, and enter **C1** into the **Averaging Terms** menu; see Figure 1.17. The resulting predictions are shown below.

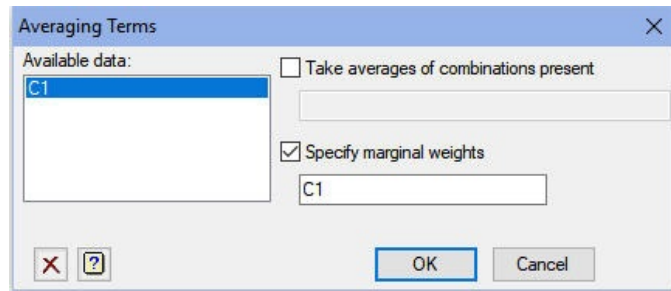


Figure 1.17

## Predictions from REML analysis

Model terms included for prediction: Constant + variety + nitrogen + variety.nitrogen

Model terms excluded for prediction: blocks + blocks.wplots

Status of model variables in prediction:

Variable	Type	Status
nitrogen	factor	Averaged over - specified weights C1
variety	factor	Classifies predictions
Constant	factor	Included in prediction
blocks	factor	Ignored
wplots	factor	Ignored

Response variate: yield

Predictions

variety	Victory	Golden rain	Marvellous
	109.4	115.7	119.8

Standard errors

variety	Victory	Golden rain	Marvellous
	8.031	8.031	8.031

Standard error of differences

variety Victory	1	*		
variety Golden rain	2	7.582	*	
variety Marvellous	3	7.582	7.582	*
		1	2	3

The menus allow you to do virtually everything but, if you prefer commands, the syntax of `VPREDICT` is given below.

---

### **VPREDICT directive**

Forms predictions from a `REML` model.

#### **Options**

<code>PRINT</code> = <i>string token</i>	What to print ( <code>description</code> , <code>predictions</code> , <code>se</code> , <code>sed</code> , <code>avesed</code> , <code>vcovariance</code> ); default <code>desc, pred, se, aves</code>
<code>CHANNEL</code> = <i>scalar</i>	Channel number for output; default * i.e. current output channel
<code>MODEL</code> = <i>formula</i>	Indicates which model terms (fixed and/or random) are to be used in forming the predictions; default * includes all the fixed terms and relevant random terms
<code>OMITTERMS</code> = <i>formula</i>	Specifies random terms to be excluded from the <code>MODEL</code> ; default * i.e. none
<code>FACTORIAL</code> = <i>scalar</i>	Limit on the number of factors or variates in each term in the models specified by <code>MODEL</code> or <code>OMITTERMS</code> ; default 3
<code>PRESENTCOMBINATIONS</code> = <i>identifiers</i>	Lists factors for which averages should be taken across combinations that are present
<code>WEIGHTS</code> = <i>tables</i>	One-way tables of weights classified by factors in the model; default *
<code>PREDICTIONS</code> = <i>table or scalar</i>	To save the predictions; default *
<code>SE</code> = <i>table or scalar</i>	To save standard errors of predictions; default *
<code>SED</code> = <i>symmetric matrix</i>	To save standard errors of differences between predictions; default *
<code>VCOVARIANCE</code> = <i>symmetric matrix</i>	To save variances and covariances of predictions; default *
<code>SAVE</code> = <i>identifier</i>	Specifies the save structure from which to predict; default * i.e. that from most recent <code>REML</code>

#### **Parameters**

<code>CLASSIFY</code> = <i>vectors</i>	Variates and/or factors to classify table of predictions
<code>LEVELS</code> = <i>variates or scalars</i>	To specify values of variates and/or levels of factors for which predictions are calculated
<code>PARALLEL</code> = <i>identifiers</i>	Specifies variables in the classifying set whose values change in parallel (rather than in all combinations)

---



The `CLASSIFY` parameter specifies those variates or factors to be included in the table of predictions, and the `LEVELS` parameter supplies the values at which the predictions are to be made. For a factor, you can select some or all of the levels, while for a variate you can specify any set of values. A single level or value is represented by a scalar; several levels or values must be combined into a variate (which may of course be unnamed). A missing value in the `LEVELS` parameter is taken to stand for all the levels of a factor, or the mean value of a variate. The `PARALLEL` parameter allows you to indicate that a factor or variate should change in parallel to another factor or variate. Both of these should have the same number of values specified for it by the `LEVELS` parameter of `VPREDICT`. The predictions are then formed for each corresponding set of values rather than for every combination of these values.

The `MODEL`, `OMITTERMS` and `FACTORIAL` options specify the model to use in step 1. The formula specified by `MODEL` is expanded into a list of model terms, deleting any that contain more variates or factors than the limit specified by the `FACTORIAL` option. Then, any random terms in the formula specified by `OMITTERMS` are removed.

The `WEIGHTS` option can supply the one-way tables to be used in step 2. These are used to calculate the weight to be used for each fitted value when calculating the averages. Equal weights are assumed for any factor for which no table of weights has been supplied. In the averaging all the fitted values are generally used. However, if you define a list of factors using the `PRESENT` option, any combination of levels of these factors that does not occur in the data will be omitted from the averaging. Where a prediction is found to be inestimable, i.e. not invariant to the model parameterization, a missing value is given.

Printed output is controlled by settings of the `PRINT` option with settings:

<code>description</code>	describes the terms and standardization policies used when forming the predictions,
<code>predictions</code>	prints the predictions,
<code>se</code>	produces predictions and standard errors,
<code>sed</code>	prints standard errors for differences between the predictions,
<code>avesed</code>	prints the average standard error of difference of the predictions, and
<code>vcovariance</code>	prints the variance and covariances of the predictions.

By default descriptions, predictions, standard errors and an average standard error of differences are printed. You can also save the results, using the `PREDICTIONS`, `SE`, `SED` and `VCOVARIANCE` options. You can send the output to another channel, or to a text structure, by setting the `CHANNEL` option.

The following statements could be used to reproduce the predictions above.

```
VPREDICT [PRINT=description,prediction,se,sed] nitrogen
VPREDICT [PRINT=description,prediction,se,sed;\
MODEL=nitrogen+variety+nitrogen.variety\
+blocks+ blocks.wplots] nitrogen
VPREDICT [PRINT=description,prediction,se,sed; WEIGHTS=C1]\
variety
```

## 1.7 Practical

Form predictions for the chemicals Practical 1.3. First of all use the default settings, and compare these with the predicted means that you obtained earlier. Then form predictions assuming that the meat will be cooked equally often at temperatures 1 and 2, but not at temperature 3.

## 1.8 A non-orthogonal design

We now consider the analysis of a rather more complicated field experiment (at Slate Hall Farm in 1976), previously analysed by Gilmour *et al.* (1995). The design was set up to study 25 varieties of wheat, and contained six replicates (each with one plot for every variety) laid out in a two by three array. The variety grown on each plot is shown in the plan below.

Each replicate has a block structure of rows crossed with columns, so the random model is

```
replicates / (rows * columns)
```

(rows crossed with columns, nested within replicates), which expands to give

```
replicates + replicates.rows + replicates.columns +  
replicates.rows.columns
```

So we have random terms for replicates, rows within replicates, columns within replicates and, finally, `replicates.rows.columns` represents the residual variation. The fixed model contains just the main effect of the factor `variety`.

1	2	4	3	5	19	23	2	6	15	18	25	9	11	2
6	7	9	8	10	8	12	16	25	4	5	7	16	23	14
21	22	24	23	25	11	20	24	3	7	6	13	22	4	20
11	12	14	13	15	22	1	10	14	18	24	1	15	17	8
16	17	19	18	20	5	9	13	17	21	12	19	3	10	21
3	18	8	13	23	16	24	10	13	2	10	4	17	11	23
1	16	6	11	21	12	20	1	9	23	12	6	24	18	5
5	20	10	15	25	4	7	18	21	15	19	13	1	25	7
2	17	7	12	22	25	3	14	17	6	21	20	8	2	14
4	19	9	14	24	8	11	22	5	19	3	22	15	9	16

The data are in spreadsheet file `Slatehall.gsh` (Figure 1.18), which can be opened using the [Example Data Sets](#) menu as shown in Figure 1.2. In addition to the factors already mentioned, the sheet also contains factors `fieldrow` and `fieldcolumn` (defining the row and column positions within the whole field, rather than within each replicate) which we shall use to define spatial correlation structures in Chapter 2.

Row	plotnumber	replicates	rows	columns	variety	yield	fieldrow	fieldcolumn
1	1	1	1	1	1	10.03	1	1
2	2	1	1	2	2	13.56	1	2
3	3	1	1	3	4	14.12	1	3
4	4	1	1	4	3	12.39	1	4
5	5	1	1	5	5	15.08	1	5
6	6	2	1	1	19	19.67	1	6
7	7	2	1	2	23	15.72	1	7
8	8	2	1	3	2	19.69	1	8
9	9	2	1	4	6	17.47	1	9
10	10	2	1	5	15	15.98	1	10
11	11	3	1	1	18	16.3	1	11
12	12	3	1	2	25	16.33	1	12
13	13	3	1	3	9	12.55	1	13
14	14	3	1	4	11	12.77	1	14
15	15	3	1	5	2	15.72	1	15

Figure 1.18

Figure 1.19 shows the [Linear Mixed Models](#) menu with the necessary boxes filled in. If we use the [Linear Mixed Model Options](#) menu (Figure 1.5) to request predicted means and standard errors of differences of means (in addition to the existing [Display options](#)), and then click on [Run](#) in the [Linear Mixed Models](#) menu itself, the following output is produced.

Figure 1.19

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + variety  
 Random model: replicates + replicates.rows + replicates.columns +  
 replicates.rows.columns  
 Number of units: 150

replicates.rows.columns used as residual term

Sparse algorithm with AI optimisation

## Estimated variance components

Random term	component	s.e.
replicates	0.4262	0.6890
replicates.rows	1.5595	0.5091
replicates.columns	1.4812	0.4865

## Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
replicates.rows.columns	Identity	Sigma2	0.806	0.1340

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	212.26	24	8.84	79.3	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	212.26	24	8.84	79.3	<0.001

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

## Table of predicted means for Constant

14.70 Standard error: 0.422

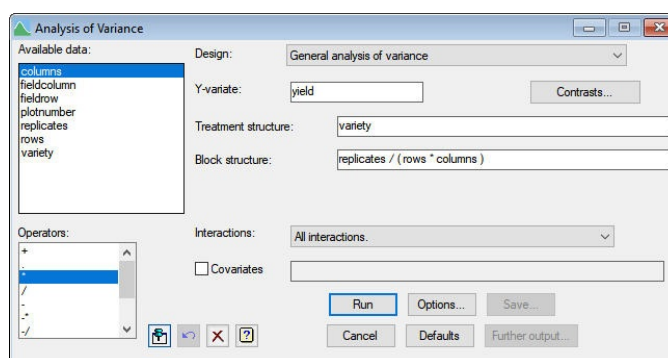
## Table of predicted means for variety

variety	1 12.84	2 15.49	3 14.21	4 14.52	5 15.33	6 15.27	7 14.01	8 14.57
variety	9 12.99	10 11.93	11 13.27	12 14.84	13 16.19	14 13.27	15 14.98	16 13.46
variety	17 14.98	18 15.92	19 16.70	20 16.40	21 14.93	22 16.44	23 13.29	24 15.46
variety	25 16.31							

Standard error of differences: 0.6202

---

Unusually for a variety trial, this particular design is balanced (in fact it is a lattice square), and we can gain additional insights into the REML analysis by looking at the output that we could have obtained from the [Analysis of Variance](#) menu. The menu is not customized for the design, but we can use the [General analysis of variance](#) setting in the [Design](#) box,



**Figure 1.20**

and specify the [Treatment structure](#) and [Block structure](#) as shown in Figure 1.20. The standard analysis of variance output (analysis-of-variance table, information summary, means and standard errors of differences) is shown below.

## Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
replicates stratum	5	133.3273	26.6655		
replicates.rows stratum variety	24	215.9053	8.9961		
replicates.columns stratum variety	24	229.8094	9.5754		
replicates.rows.columns stratum variety	24	166.7675	6.9486	8.58	<.001
Residual	72	58.3011	0.8097		
Total	149	804.1105			

## Information summary

Model term	e.f.	non-orthogonal terms
replicates.rows stratum variety	0.167	
replicates.columns stratum variety	0.167	replicates.rows
replicates.rows.columns stratum variety	0.667	replicates.rows replicates.columns

*Message: the following units have large residuals.*

replicates 6	-1.895	approx. s.e. 0.943
replicates 1 rows 4 columns 3	-1.665	approx. s.e. 0.623

replicates 1 rows 5 columns 2

1.710

approx. s.e. 0.623

## Tables of means

Variate: yield

Grand mean 14.704

variety	1	2	3	4	5	6	7
	12.962	15.561	14.152	14.560	15.481	15.358	14.008
variety	8	9	10	11	12	13	14
	14.428	12.968	11.928	13.222	14.835	16.176	13.187
variety	15	16	17	18	19	20	21
	15.067	13.287	14.968	15.881	16.742	16.277	15.048
variety	22	23	24	25			
	16.430	13.283	15.464	16.344			

## Standard errors of differences of means

Table	variety
rep.	6
d.f.	72
s.e.d.	0.6363

---

Notice that the analysis-of-variance table has *three* lines for `variety`. As each row contains a different set of varieties, the differences between the rows in each replicate enable us to obtain estimates of the variety effects (which appear in the `replicates.rows` stratum). The same is true of the columns. The design is balanced because the various comparisons between varieties are all estimated with the same efficiency in the `replicates.rows` stratum; the Information Summary indicates the efficiency is in fact 0.167. Similarly, they all have efficiency 0.167 in the `replicates.columns` stratum, and efficiency 0.667 in the `replicates.rows.columns` stratum. So, the possible information on variety is split ( $1/6 : 1/6 : 2/3$ ) between the three strata.

We can see the estimates obtained in each stratum by checking the **Effects** box in the **ANOVA Further Output** menu (Figure 1.21) and clicking on **Run**. You can then verify that the standard table of means produced by **ANOVA**, above, is calculated using the estimated effects from the lowest stratum (`replicates.rows.columns`): the mean 12.962 for variety 1 is the grand mean 14.704 plus the effect of variety 1 in the `replicates.rows.columns` table, namely  $-1.742$ .

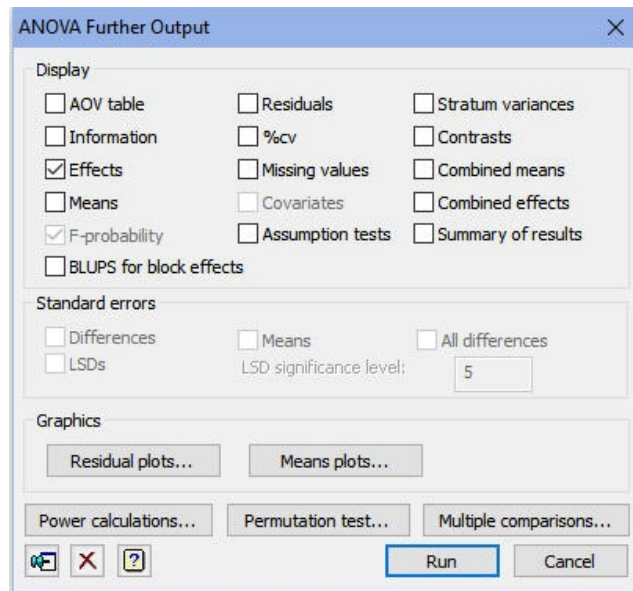


Figure 1.21

## Tables of effects

Variate: yield

### `replicates.rows` stratum

variety effects, e.s.e. \*, rep. 6

variety	1	2	3	4	5	6	7
	-5.614	1.296	0.604	-1.468	-3.522	2.790	-3.458
variety	8	9	10	11	12	13	14
	1.718	0.520	-3.814	-2.718	-2.544	1.020	1.236
variety	15	16	17	18	19	20	21
	0.582	5.598	3.786	3.480	3.902	3.530	-1.294
variety	22	23	24	25			
	-0.028	1.360	-3.058	-3.894			

### `replicates.columns` stratum

variety effects, e.s.e. \*, rep. 6

variety	1	2	3	4	5	6	7
	-3.432	-2.588	0.812	-0.650	-1.450	-4.948	1.930
variety	8	9	10	11	12	13	14
	4.064	-3.010	-1.584	1.852	2.828	2.540	-0.752
variety	15	16	17	18	19	20	21
	-3.536	-0.642	-2.494	0.740	-1.706	4.934	-2.924



variety	22	23	24	25
	3.990	-3.730	4.434	5.332

### replicates.rows.columns stratum

variety effects, e.s.e. 0.4499, rep. 6

variety	1	2	3	4	5	6	7
	-1.742	0.857	-0.553	-0.144	0.777	0.653	-0.697
variety	8	9	10	11	12	13	14
	-0.277	-1.736	-2.777	-1.482	0.130	1.471	-1.517
variety	15	16	17	18	19	20	21
	0.362	-1.418	0.263	1.176	2.037	1.573	0.343
variety	22	23	24	25			
	1.726	-1.421	0.760	1.639			

In contrast, the [REML](#) analysis has produced a single set of estimates, and these automatically combine (with an appropriate weighting) all the separate estimates. In fact the [REML](#) estimates correspond to the *combined* effects and means in the [ANOVA Further Output](#) menu. Below, we show these tables, together with the output generated by checking the [Stratum variances](#) box which contains the variance components. The combined means have a smaller standard error of difference than the standard means, but the complicated structure of their estimation means that we can no longer assume that differences between them follow t-distributions with a known number of degrees of freedom. (However, the *effective* numbers of degrees of freedom printed by [ANOVA](#) are generally reasonably reliable.)

## Tables of combined effects

Variate: yield

variety effects, e.s.e. 0.4385, rep. 6, effective d.f. 79.99

variety	1	2	3	4	5	6	7
	-1.869	0.786	-0.495	-0.186	0.628	0.570	-0.697
variety	8	9	10	11	12	13	14
	-0.131	-1.716	-2.772	-1.432	0.133	1.486	-1.438
variety	15	16	17	18	19	20	21
	0.276	-1.243	0.277	1.217	1.991	1.695	0.230
variety	22	23	24	25			
	1.739	-1.413	0.760	1.602			

## Tables of combined means

Variate: yield

variety	1	2	3	4	5	6	7
	12.836	15.490	14.209	14.519	15.333	15.274	14.007
variety	8	9	10	11	12	13	14
	14.574	12.989	11.932	13.272	14.838	16.190	13.266
variety	15	16	17	18	19	20	21
	14.980	13.461	14.982	15.922	16.696	16.399	14.934
variety	22	23	24	25			
	16.444	13.291	15.465	16.306			

## Standard errors of differences of combined means

Table	variety
rep.	6
s.e.d.	0.6202
effective d.f.	79.99

## Estimated stratum variances

Variate: yield

Stratum	variance	effective d.f.	variance component
replicates	26.6655	5.000	0.4262
replicates.rows	8.6037	23.464	1.5595
replicates.columns	8.2120	23.438	1.4812
replicates.rows.columns	0.8062	73.099	0.8062

The example reinforces the point that the **REML** output is the same as that given by **ANOVA** when both are feasible, but that the generality of the **REML** method leaves aspects that it cannot duplicate. More importantly, though, it shows that the **REML** method makes use of all the available information about each fixed effect. These aspects indicate the efficiency and appropriateness of the methodology, and the next practical illustrates its ability to handle designs that cannot be analysed by **ANOVA**.

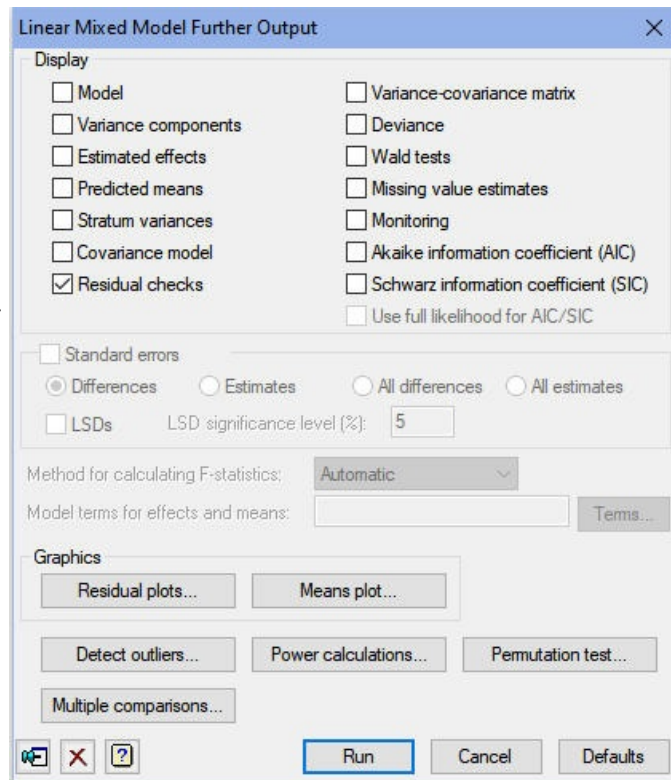
## 1.9 Practical

Genstat spreadsheet file **Vartrial1.gsh** contains data from a trial of 35 varieties of wheat. The design has two replicates each laid out in a five by seven plot array. Assuming that the same block structure is appropriate as in Section 1.8 (rows crossed with columns within replicates), analyse the data as a linear mixed model with block structure **Replicate/(Row\*Column)** and fixed model **Variety**. Check whether the data can be analysed by **ANOVA**.

## 1.10 Residual checks and plots

You can ask Genstat to report any large residuals, by checking the [Residual Checks](#) box on either the [Linear Mixed Model Options](#) menu or the [Linear Mixed Model Further Output](#) menu, as in Figure 1.22.

The criterion for deciding whether a standardized residual is large depends on the number of degrees of freedom  $d$  of the random terms in the analysis. Genstat reports any residual with absolute value greater than 2 if  $d$  is less than 20, or greater than 4.0 if  $d$  is greater than 15773. For other values of  $d$ , the default is the critical value of the Normal distribution for a two-sided test with significance probability  $1/d$ . These criteria are the same as those used in regression and analysis of variance, and are



**Figure 1.22**

intended to ensure that a report should appear for any extreme outlier, but that reports should not appear too often just as a result of random variation.

For the Slate Hall data, two large residuals are reported, in units 48 and 62. It would be worthwhile checking the original records to see if there was anything unusual about those plots, perhaps to justify their exclusion from the analysis.

## Large residuals

---

Unit	Residual
48	-2.747
62	3.227

---

You can use the [REML Residual Plot Options](#) menu (Figure 1.23) to produce diagnostic plots of the residuals, so that you can check the assumptions of the analysis. Alternatively, if you are analysing a field experiment, you can display the residuals in field layout to check for systematic trends up and down or across the field. The [Method](#) list box controls whether the display shows just the residuals from the final random term (to check the assumptions), or the sum of all the random effects (to assess fertility trends).

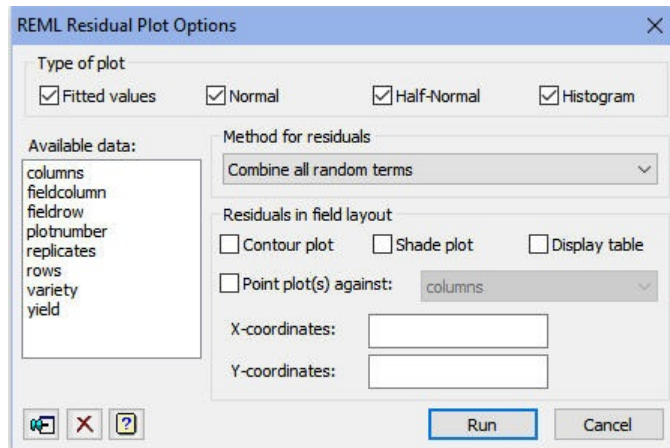


Figure 1.23

To open the menu you click on the [Further output](#) button in the [Linear Mixed Models](#) menu (Figure 1.22), and then clicking on the [Residual plots](#) button in the [Linear Mixed Model Further Output](#) menu (Figure 1.6).

The menu has check boxes to provide four types of diagnostic plot. In Figure 1.23 they are all checked, so Figure 1.24 shows them all. There is a histogram of the residuals, so that you can check that the distribution is symmetrical and reasonably Normal; a plot of residuals against fitted values, so that you can check whether the residuals are roughly symmetrically distributed with constant variance; a Normal plot which

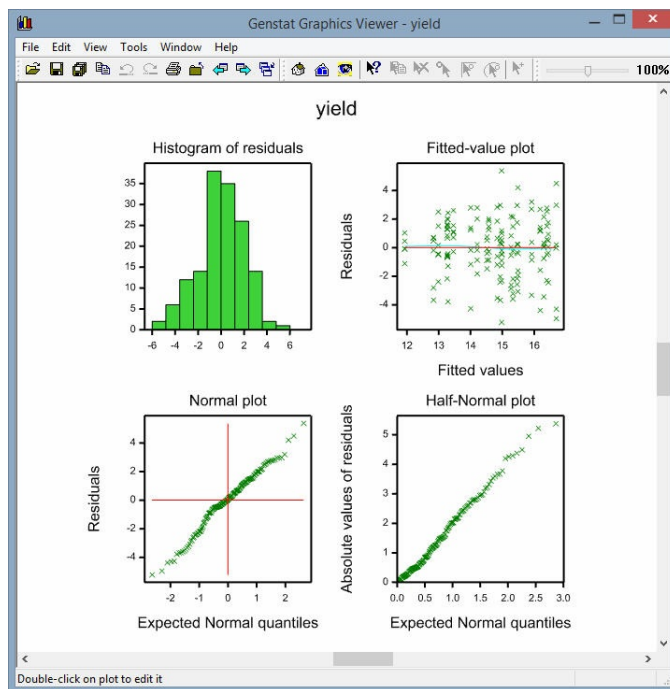


Figure 1.24

plots the ordered residuals against Normal distribution statistics – if they lie roughly on a straight line, the residuals are roughly Normally distributed; and a half-Normal plot which does the same for the absolute values of the residuals, and can be more useful for small sets of data. The same plots are available from the analysis of variance menus; see Chapter 4 of the *Guide to Anova and Design in Genstat* for more details.

The diagnostic plots are produced by the `VPLOT` procedure. If you prefer commands to menus, the syntax is described below.

**VPLOT procedure**

Plots residuals from a REML analysis (S.J. Welham).

**Options**

<code>RMETHOD</code> = <i>string token</i>	Which random terms to use when calculating the residuals ( <code>final</code> , <code>all</code> , <code>notspline</code> , <code>stfinal</code> , <code>stall</code> ); default uses the setting from the <code>REML</code> statement
<code>INDEX</code> = <i>variate or factor</i>	X-variable for an index plot; default <code>!(1,2...)</code>
<code>GRAPHICS</code> = <i>string token</i>	What type of graphics to use ( <code>lineprinter</code> , <code>highresolution</code> ); default <code>high</code>
<code>TITLE</code> = <i>text</i>	Overall title for the plots; if unset, the identifier of the y-variate is used
<code>SAVE</code> = <i>REML save structure</i>	Specifies the ( <code>REML</code> ) save structure from which the residuals and fitted values are to be taken; default <code>*</code> uses the <code>SAVE</code> structure from the most recent <code>REML</code> analysis

**Parameters**

<code>METHOD</code> = <i>string tokens</i>	Type of residual plot ( <code>fittedvalues</code> , <code>normal</code> , <code>halfnormal</code> , <code>histogram</code> , <code>absresidual</code> , <code>index</code> ); default <code>fitt</code> , <code>norm</code> , <code>half</code> , <code>hist</code>
<code>PEN</code> = <i>scalars, variates or factors</i>	Pen(s) to use for each plot

Procedure `VPLOT` provides up to four types of residual plots from a `REML` analysis. These are selected using the `METHOD` parameter, with settings: `fitted` for residuals versus fitted values, `normal` for a Normal plot, `halfnormal` for a half-Normal plot, `histogram` for a histogram of residuals, `absresidual` for a plot of the absolute values of the residuals versus the fitted values, and `index` for a plot against an "index" variable (specified by the `INDEX` option). The `PEN` parameter can specify the graphics pen or pens to use for each plot. The `TITLE` option can supply an overall title. If this is not set, the identifier of the y-variate is used.

The residuals and fitted values are accessed automatically from the analysis specified by the `SAVE` option. If the `SAVE` option has not been set, they are taken from the `SAVE` structure from the most recent `REML` analysis.

The `RMETHOD` option controls which random terms are used to calculate the residuals:

<code>all</code>	all the random effects,
<code>final</code>	only the final random term,
<code>notspline</code>	all except any random spline terms,
<code>stall</code>	standardized residuals using all the random effects, and
<code>stfinal</code>	standardized residuals using only the final random term.

The default takes the setting from the `REML` directive that produced the analysis. Note that residuals based on the final random term will not be calculated when any of the variance

components are negative, as the associated negative correlations can generate very misleading patterns. `V PLOT` will then generate a warning that all the residuals are missing, and you should use `RMETHOD=all` instead.

By default, high-resolution graphics are used. Line-printer graphics can be used by setting option `GRAPHICS=lineprinter`.

The plots of the residuals in field layout are produced by the `VDFIELDRESIDUALS` procedure.

## 1.11 Practical

Try to plot the residuals from the `REML` analysis of the data in spreadsheet file `Vartrial1.gsh`. You will find that you are unsuccessful, as Genstat does not allow you to save residuals for the final error term when there are negative variance components. (This is because these random terms can generate strange correlation patterns between the residuals that can be confusing and hard to interpret.) Remove the random terms with the negative variance components, by redefining the random model as `Replicate.Column`, and try again. (The term `Replicate.Column` now fits a random effect for every combination of the `Replicate` and `Column` factors i.e. for every physical column in the design.)

## 1.12 Saving information from the analysis

As well as displaying the results of an analysis, the `REML` menus allow you to save the results in standard data structures. After a `REML` analysis you can click on the `Save` button of the `Linear Mixed Models` menu (Figure 1.4), to open the `Linear Mixed Models Save Options` menu. The residuals, fitted values, predicted means and many other results can be saved. In Figure 1.25 `variety` is selected as the model term for which means and effects are to be saved, the `Predicted Means` box is checked, and `varmeans` is entered into the adjacent box as the name of the table to store them. The information is saved using the `VKEEP` directive.

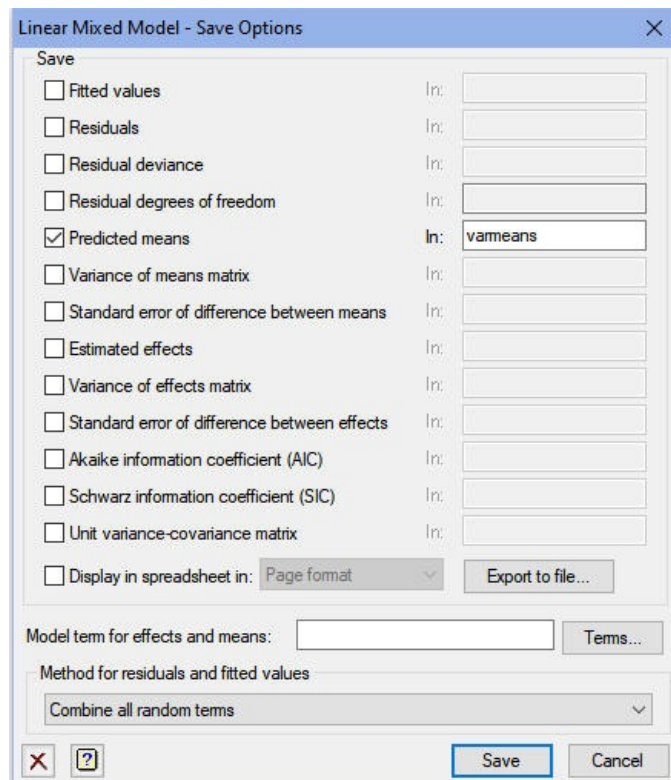


Figure 1.25



If you check the [Display in spreadsheet](#) box, the results are put into a Genstat spreadsheet, which can then be saved in a file on your computer for use in a later run of Genstat, or in another program such as Excel. Alternatively you can save results automatically to a spreadsheet file by clicking on the [Export to file](#) button. This opens the [Save REML Results in Spreadsheet File](#) menu. Figure 1.26, shows the menu with the default output components selected in the check boxes, and the [Save in file](#) box filled in to save them in the Excel file [SlateHallResults.xlsx](#).

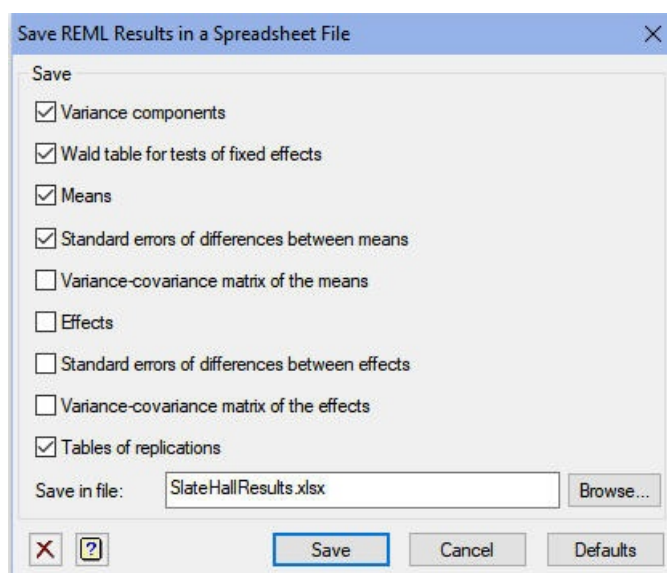


Figure 1.26

Each section of output is saved in a separate page of the spreadsheet file. Figure 1.27 shows the page containing the variance components.

The menu uses the `VSPREADSHEET` procedure.

	A	B	C	D	E	F	G	H	I	J	K	L
1	UnitLabels	components										
2	replicates	0.426238622										
3	replicates.rows	1.559506138										
4	replicates.columns	1.48115487										

Figure 1.27



---

## 2 Meta analysis with REML

In this chapter we describe how you can use **REML** to do a combined analysis of several related experiments. The aim here is to produce estimates of treatment effects that make use of all the available information. This form of *meta analysis* gives the most efficient estimates, provided all the original data sets are still available. If the original data sets are no longer available, and you have only the results from the analysis of each one, you can use the **META** procedure (or its associated menu [Meta Analysis of Trial Results](#)) for a single treatment contrast, or the **VMETA** procedure (or its associated menu [Multi-treatment Meta Analysis of Summaries](#)) for several treatment effects.

However, in order for the combined analysis to be sensible, the experiments should have similar treatment structures, and should have some treatments in common across the experiments. Otherwise, there is no information on comparisons between pairs of treatments that are not in the same experiment.

So, in this chapter you will learn

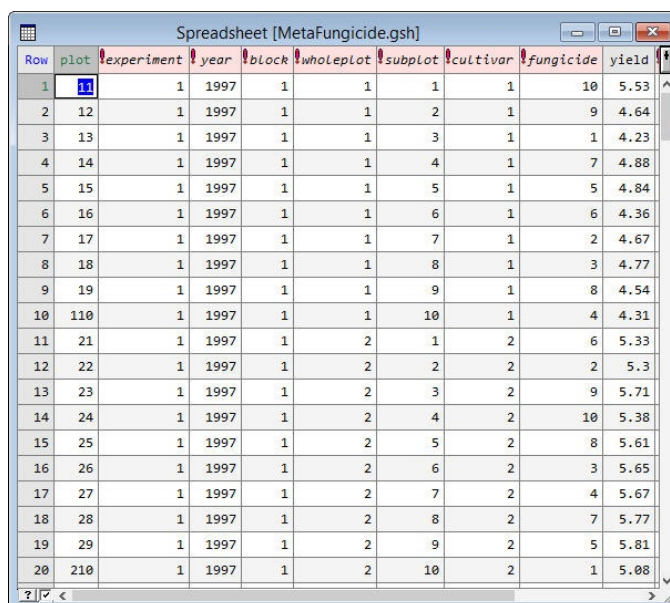
- how to use the [Multiple Experiments / Meta Analysis \(REML\)](#) menu
- the **VRESIDUAL** command★

Note: the topics marked ★ are optional.

## 2.1 Example: a series of fungicide trials

As an example we shall consider the analysis of three fungicide trials that took place in different years at the same site. The data are in spreadsheet file *MetaFungicide.gsh* (Figure 2.1). There were two cultivars, one susceptible and one resistant, and ten different fungicide treatments. A split-plot design was used in each year, but the cultivars were applied to the whole-plots in 1997, and the fungicides were applied to the whole-plots in 1998 and 1999. So, we have the same treatments, but different designs in the different years, even though the blocking structures were identical.

We can analyse the individual trials by using the spreadsheet menus to restrict the data set to experiments 1, 2 and 3 in turn. We first open the **Restrict Units on Factor** menu, by selecting the **To Groups (Factor levels)** sub-option of the **Restrict/Filter** option of the **Spread** menu on the menu bar (see Figure 2.2).



Row	plot	experiment	year	block	wholeplot	subplot	cultivar	fungicide	yield
1	11	1	1997	1	1	1	1	10	5.53
2	12	1	1997	1	1	2	1	9	4.64
3	13	1	1997	1	1	3	1	1	4.23
4	14	1	1997	1	1	4	1	7	4.88
5	15	1	1997	1	1	5	1	5	4.84
6	16	1	1997	1	1	6	1	6	4.36
7	17	1	1997	1	1	7	1	2	4.67
8	18	1	1997	1	1	8	1	3	4.77
9	19	1	1997	1	1	9	1	8	4.54
10	110	1	1997	1	1	10	1	4	4.31
11	21	1	1997	1	2	1	2	6	5.33
12	22	1	1997	1	2	2	2	2	5.3
13	23	1	1997	1	2	3	2	9	5.71
14	24	1	1997	1	2	4	2	10	5.38
15	25	1	1997	1	2	5	2	8	5.61
16	26	1	1997	1	2	6	2	3	5.65
17	27	1	1997	1	2	7	2	4	5.67
18	28	1	1997	1	2	8	2	7	5.77
19	29	1	1997	1	2	9	2	5	5.81
20	210	1	1997	1	2	10	2	1	5.08

Figure 2.1

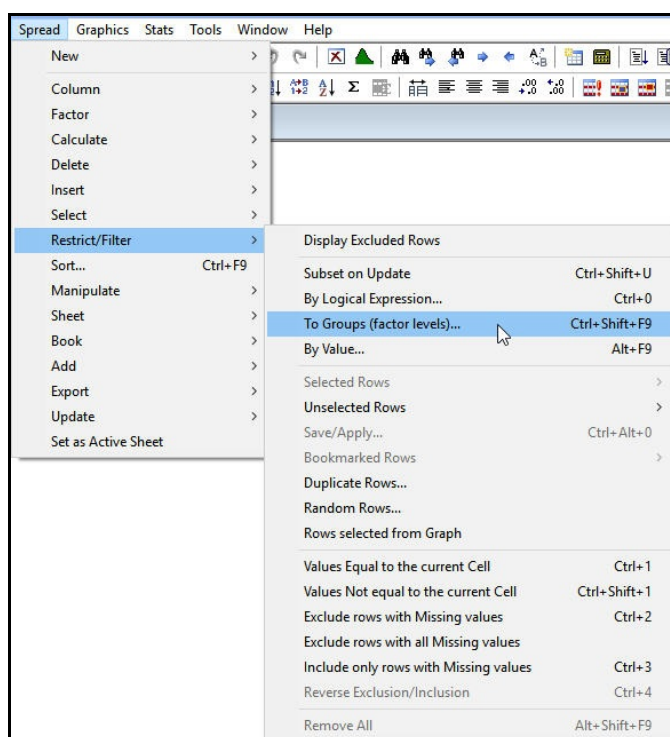


Figure 2.2

Then, in that menu, we select the experiment (1, 2 or 3 in turn) and click on **OK**. The same menu should be used to remove the restriction after the analyses.

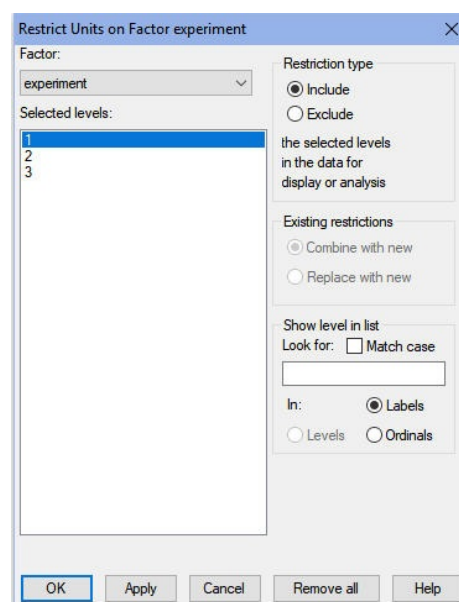


Figure 2.3

The analysis is specified in the **Linear Mixed Models** menu, similarly to the split-plot analysis in Section 1.2. The settings are shown in Figure 2.4, and the output from analysing each experiment, in turn, is given below.

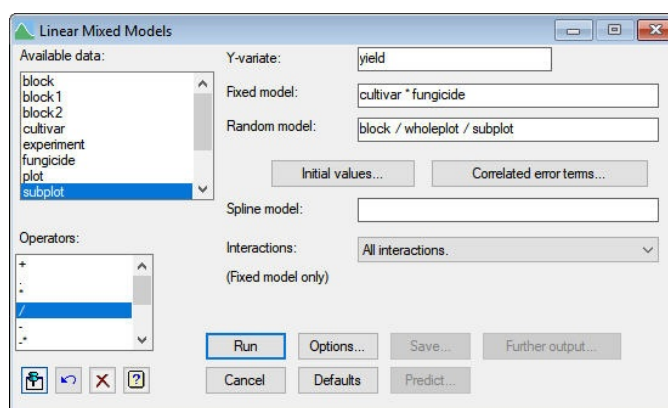


Figure 2.4

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + cultivar + fungicide + cultivar.fungicide  
 Random model: block + block.wholeplot + block.wholeplot.subplot  
 Number of units: 60

block.wholeplot.subplot used as residual term

Sparse algorithm with AI optimisation  
 Analysis is subject to the restriction on yield

## Estimated variance components

Random term	component	s.e.
block	0.01354	0.01503
block.wholeplot	-0.00165	0.00303

## Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
block.wholeplot.subplot	Identity	Sigma2	0.0449	0.01059

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar	311.24	1	311.24	2.0	0.003
fungicide	60.02	9	6.67	36.0	<0.001
cultivar.fungicide	37.98	9	4.22	36.0	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar.fungicide	37.98	9	4.22	36.0	<0.001

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + cultivar + fungicide + cultivar.fungicide  
 Random model: block + block.wholeplot + block.wholeplot.subplot  
 Number of units: 60

block.wholeplot.subplot used as residual term

Sparse algorithm with AI optimisation  
 Analysis is subject to the restriction on yield

## Estimated variance components

Random term	component	s.e.
block	0.01512	0.01657
block.wholeplot	-0.02177	0.01241

## Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
block.wholeplot.subplot	Identity	Sigma2	0.0724	0.02289

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar	45.26	1	45.26	20.0	<0.001
fungicide	378.50	9	42.06	18.0	<0.001
cultivar.fungicide	12.12	9	1.35	20.0	0.275

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar.fungicide	12.12	9	1.35	20.0	0.275

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + cultivar + fungicide + cultivar.fungicide  
 Random model: block + block.wholeplot + block.wholeplot.subplot  
 Number of units: 60

block.wholeplot.subplot used as residual term

Sparse algorithm with AI optimisation  
 Analysis is subject to the restriction on yield

## Estimated variance components

Random term	component	s.e.
block	-0.0043	0.0038
block.wholeplot	0.0084	0.0317

## Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
block.wholeplot.subplot	Identity	Sigma2	0.129	0.0407

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar	37.41	1	37.41	20.0	<0.001
fungicide	102.08	9	11.34	18.0	<0.001
cultivar.fungicide	11.53	9	1.28	20.0	0.306

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar.fungicide	11.53	9	1.28	20.0	0.306

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

The variance components can be tabulated as follows:

Year	Blocks	Whole-plots	Residual
1997	0.01354	-0.00165	0.0449
1998	0.01512	-0.02177	0.0724
1999	-0.0043	0.0084	0.129

The F-probabilities for treatment terms show an interaction only in 1997.

Year	Cultivar	Fungicide	Interaction
1997	0.003	<.001	<.001
1998	<.001	<.001	0.275
1999	<.001	<.001	0.306

To illustrate how you specify different error models for each experiment, we will omit the negative variance components, i.e. treat them as zero. (In practice we could study this further during the meta analysis.) So for each of experiments 1 and 2 (1997 and 1998) we need a random term for blocks, while for experiment 3 (1999) we need a random term for the combinations of whole-plots and blocks. These require extra factors which can be set up by the commands

```
CALCULATE  block1 = MVINSERT(block; experiment.NE.1)
&          block2 = MVINSERT(block; experiment.NE.2)
&          wplot3 = MVINSERT(wholeplot; experiment.NE.3)
GROUP      [REDEFINE=yes] block1,block2,wplot3
```

or by using the [Calculate](#) and [Form Groups](#) menus (accessed by selecting the [Calculations](#) and [Form Groups \(Factors\)](#) options of the [Data](#) menu on the menu bar). The results can be found in the right-hand columns of the spreadsheet file [MetaFungicide.gsh](#). Notice that each of these factors has missing values except in the units belonging to its own experiment. Another possibility is to use the [VRMETAMODEL](#) procedure. Details can be found in Section 5.8.1 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

We can now represent the random terms arising from the blocking structures by the model formula

```
block1 + block2 + block.wplot3
```

In addition, the meta analysis menu will automatically estimate the residual variance separately within each experiment.

We also need to consider how to handle experiment effects and interactions between experiments and the treatment terms. If we include these as in the fixed model, the treatment terms will be tested using the within-experiment error, weighted according to precision within each experiment. Alternatively, if we include them in the random model, each treatment term will in effect be compared with its interaction with experiment (unless this is zero). In that case, a significant treatment effect would imply that the effect is consistent and large compared to its variation across experiments – thus giving a more stringent test.

We now open the **Multiple Experiments/Meta Analysis (REML)** menu, by selecting the **Multiple Experiments/Meta Analysis** sub-option of the **REML** option of the **Stats** menu on the menu bar.

The fixed model is `cultivar*fungicide`, as in the individual analyses, and the random model is

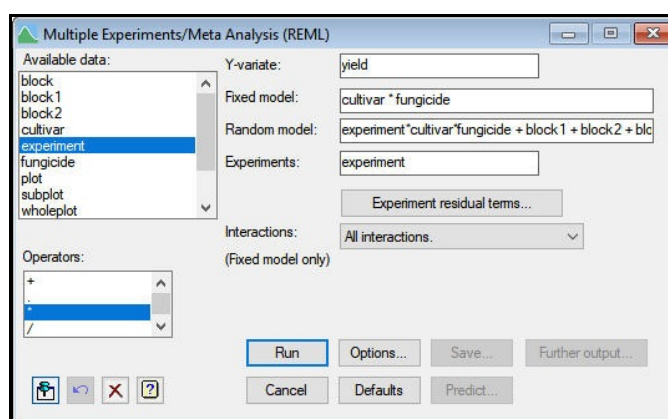


Figure 2.5

```
experiment*cultivar*fungicide + block1 + block2 + block.wplot3
```

as explained above. (Note that we can use the simple form `experiment*cultivar*fungicide` to specify the experiment-by-treatment interactions as any terms that occur in both the fixed and the random models for REML are dropped from the random model.) The y-variate is again `yield`, and the factor identifying the experiments is `experiment`. The output is shown below.

## REML variance components analysis

Response variate:	yield
Fixed model:	Constant + cultivar + fungicide + cultivar.fungicide
Random model:	experiment + experiment.cultivar + experiment.fungicide + experiment.cultivar.fungicide + block1 + block2 + block.wplot3
Number of units:	180

Separate residual terms for each level of experiment factor: experiment

Sparse algorithm with AI optimisation

Units with missing factor/covariate values included

- specific effect for term(s) omitted for units with missing values in block1, block2, wplot3



## Estimated variance components

Random term	component	s.e.
experiment	0.5984	0.6107
experiment.cultivar	0.0094	0.0127
experiment.fungicide	0.0223	0.0147
experiment.cultivar.fungicide	0.0127	0.0114
block1	0.0126	0.0150
block2	0.0141	0.0166
block.wplot3	0.0087	0.0271

## Residual model for each experiment

Experiment factor: experiment

Experiment	Term	Model(order)	Parameter	Estimate	s.e.
1	Residual	Identity	Variance	0.0472	0.0112
2	Residual	Identity	Variance	0.0494	0.0110
3	Residual	Identity	Variance	0.118	0.034

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar	41.92	1	41.92	2.1	0.021
fungicide	92.10	9	10.23	17.2	<0.001
cultivar.fungicide	22.61	9	2.51	16.5	0.050

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
cultivar.fungicide	22.61	9	2.51	16.5	0.050

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

---

The estimated variance components for the blocking terms and residuals are similar to those from individual experiments (as you might expect). The variance components for the random terms involving experiment are small except for the experiment main effects. The main effects of cultivar and fungicide are still significant across the three experiments, and there is some evidence of an interaction.

## 2.2 Commands for meta analysis

The commands that are defined to produce the meta analysis in Section 2.1 are a straightforward extension of those in Chapter 1.

```
VCOMPONENTS [FIXED=cultivar*fungicide; FACTORIAL=9; \
  EXPERIMENTS=experiment]\
  RANDOM=experiment*fungicide*cultivar\
  + block1 + block2 + block.wplot3
REML [PRINT=model,components,waldTests;\
  MVINCLUDE=explanatory] yield
```

Notice that the `EXPERIMENTS` option is used to define the experiment factor (a different residual variance is then estimated within each of its levels). Then option `MVINCLUDE=explanatory` is set to ensure that units with missing values in any of the explanatory variables are still included in the analysis. In the analysis above, this involves the factors `block1`, `block2` and `wplot3`. These then make no contribution to the units where they are missing. So different random terms are then fitted with each of the experiments.

The [Experiment Residual Terms](#) button of the [Multiple Experiments/Meta Analysis \(REML\)](#) menu allows you to specify a different residual term for an experiment, or you can define a correlation model for the residual term. This is done using the `VRESIDUAL` directive. We will not illustrate this in the course, but the details are given below for future reference. Correlation models are described in Chapter 3.

---

### VRESIDUAL directive

Defines the residual term for a `REML` analysis, or the residual term for an experiment within a meta-analysis (combined analysis of several experiments).

#### Options

<code>EXPERIMENT</code> = <i>scalar</i>	Level of the <code>EXPERIMENTS</code> factor for which the residual is being defined
<code>TERM</code> = <i>formula</i>	Model term to be used as the residual
<code>FORMATION</code> = <i>string token</i>	Whether the structure is formed by direct product construction or by definition of the whole matrix ( <code>direct</code> , <code>whole</code> ); default <code>dire</code>
<code>VARIANCE</code> = <i>scalar</i>	Allows an initial estimate to be provided for the residual variance of the experiment
<code>CONSTRAINT</code> = <i>string token</i>	Allows the residual variance to be fixed at its initial value ( <code>fix</code> , <code>positive</code> ) default <code>posi</code>
<code>COORDINATES</code> = <i>matrix or variates</i>	Coordinates of the data points to be used in calculating distance-based models

#### Parameters

<code>MODELTYPE</code> = <i>string tokens</i>	Type of covariance model associated with the term(s), or with individual factors in the term(s) if <code>FORMATION=direct</code> ( <code>identity</code> , <code>fixed</code> , <code>AR</code> , <code>MA</code> , <code>ARMA</code> , <code>power</code> , <code>boundedlinear</code> , <code>circular</code> , <code>spherical</code> , <code>linearvariance</code> , <code>banded</code> , <code>correlation</code> , <code>antedependence</code> ,
---	---

	unstructured, diagonal, uniform, FA, FAequal) default iden
ORDER = scalar	Order of model
HETEROGENEITY = string token	Heterogeneity for correlation matrices (none, outside); default none
METRIC = string token	How to calculate distances when MODELTYPE=power (cityblock, squared, euclidean); default city
FACTOR = factors	Factors over which to form direct products
MATRIX = identifiers	To define matrix values for the term or the factors when MODELTYPE=fixed
INVERSE = identifiers	To define values for matrix inverses (instead of the fixed matrices themselves) when MODELTYPE=fixed
INITIAL = identifiers	Initial parameter values for each correlation matrix
CONSTRAINTS = texts	Texts containing strings none, fix or positive to define constraints for the parameters in each model
EQUALITYCONSTRAINTS = variates	Non-zero values in the variate indicate groups of parameters whose values are to be constrained to be equal

---

VRESIDUAL is used to define the residual term for a REML analysis or to define separate residual terms for different experiments within a meta-analysis. (For a single experiment, VRESIDUAL can be used to impose a covariance structure on the residual term. This could also be done by specifying the covariance structure using VSTRUCTURE, as explained in Chapter 3, but VRESIDUAL has the advantage that the algorithm then checks that the term is consistent with the structure of the data.)

The TERM option is used to specify the formula for the residual term. This term need not have been specified previously by the VCOMPONENTS statement.

The EXPERIMENT option is used to specify the experiment(s) for which the model is to be used. The settings identify levels of a factor, defining the experiments, which is specified by the EXPERIMENTS option of VCOMPONENTS.

The VARIANCE option is used to give an initial value for the residual variance in the current experiment(s). You can set option CONSTRAINT=fix to fix the residual variance at the initial value rather than estimating it (as a positive value).

The definition of the residual terms then follows mainly as for the definition of correlated error terms through VSTRUCTURE. The exception is that power models can be defined only in terms of the coordinates of the data points, not by specifying coordinates for the factor levels. (See Section 3.3.)

The factors and variates for the separate experiments should be concatenated into structures which run over all the experiments. When some factors differ between experiments, these should be defined on the units relevant to the appropriate experiment(s) and missing elsewhere. When an EXPERIMENTS factor has been defined, the default action of the MVINCLUDE option of REML is changed to include units with missing y-values and missing factor levels.

## 2.3 Practical

Genstat spreadsheet file `Recovery.gsh` contains data from trials of two anaesthetic agents (A and B) at nine different centres (see Whitehead 2002, *Meta-Analysis of Controlled Clinical Trials*, Section 3.6.1). The patients had undergone short surgical procedures, where a rapid recovery was regarded as important. The analysis variable is thus the log of the recovery time of each patient.

Perform a meta analysis using all the data, and including centre and the centre-by-anaesthetic interaction as random terms.

---

## 3 Spatial analysis

In this chapter we show how **REML** can model spatial correlations between observations in two-dimensions. These methods have proved very successful, for example, in the analysis of field experiments to assess new plant varieties, where the designs usually contain too many varieties for the conventional blocking techniques to be effective.

So you will learn

- how to model covariances between effects of a random term
- how to assess different covariance models
- the advantages of representing the variation of a 2-dimensional experiment by modelling its spatial covariances, as compared to using conventional blocking
- the **VSTRUCTURE** command ★

Note: the topics marked ★ are optional.

### 3.1 Traditional blocking

Traditional field experiments use blocking to improve the precision with which the treatment effects are estimated. For example, in a complete-randomized-block design, units are grouped into *blocks*, so that

- units in the same block are more similar than units in different blocks,
- each block contains the same number of reps (usually 1) of each treatment combination,
- allocation of the treatments is randomized independently within each block

The analysis estimates & removes between-block differences into the block stratum so that treatment effects can be estimated more precisely.

The output below shows the analysis that would be obtained if the field experiment at Slate Hall Farm (Section 1.8) was analysed as a randomized block design with *replicates* as the block factor.

#### Analysis of variance

Variate: yield

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
replicates stratum	5	133.327	26.665	7.69	
replicates.*Units* stratum					
variety	24	254.808	10.617	3.06	<.001
Residual	120	415.976	3.466		
Total	149	804.110			

*Message: the following units have large residuals.*

replicates 6	-1.90	s.e. 0.94
replicates 1 *units* 22	5.19	s.e. 1.67
replicates 1 *units* 23	4.69	s.e. 1.67
replicates 5 *units* 15	-4.46	s.e. 1.67
replicates 5 *units* 25	-4.45	s.e. 1.67

#### Tables of means

Variate: yield

Grand mean 14.70

variety	1	2	3	4	5	6	7
	12.04	15.06	14.57	14.26	14.39	14.78	13.99
variety	8	9	10	11	12	13	14
	15.48	13.13	11.95	13.57	14.84	16.28	13.77

variety	15	16	17	18	19	20	21
	14.45	14.59	15.10	16.19	16.43	17.16	14.23
variety	22	23	24	25			
	16.52	13.36	15.44	16.04			

## Standard errors of differences of means

Table	variety
rep.	6
d.f.	120
s.e.d.	1.075

---

The variety sum of squares is still significant, but the standard error of difference for the means (1.075) is larger than that from the analysis of variance of the design as a lattice square (0.6363) in Section 1.4 – showing that the randomized-block analysis has not modelled the patterns of fertility in the field particularly well.

In fact, the fertility is modelled here by a single block effect, which generates a uniform correlation between the plots within each block. To look at how this represents the fertility in the field, we need to examine the *combined* residuals from the analysis of variance (that is, the usual residuals from the `replicates.*Units*` stratum, plus the replicate effects). The analysis of variance save menus will only save the usual residuals, but we can use the `CBRESIDUALS` option of the `AKEEP` directive. The program below does the analysis of variance, saves the combined residuals and the replicate effects, and plots them against columns for the upper replicates (1-3) and then for the lower replicates (4-6).

```
BLOCKS replicates
TREATMENTS variety
ANOVA [PRINT=aovtable,information,means; FPROB=yes] yield
"plot the replicate effects and the combined residuals"
AKEEP [CBRESIDUALS=cbresiduals] replicates; RESIDUALS=reptab
VARIATE repvar; VALUES=reptab
CALCULATE reptrend = NEWLEVELS(replicates; repvar)
PEN 1...6; COLOUR='black','red','green'; SYMBOL=1,2,3
PEN 11; COLOUR='black'; METHOD=line; SYMBOL=0
RESTRICT cbresiduals,reptrend; replicates .IN. !(1,2,3)
DGRAPH cbresiduals,reptrend; fieldcolumn; PEN=replicates,11
RESTRICT cbresiduals,reptrend; replicates .IN. !(4,5,6)
DGRAPH cbresiduals,reptrend; fieldcolumn; PEN=replicates,11
RESTRICT cbresiduals,reptrend
```

The plots, in Figures 3.1 and 3.2, reinforce the point that the fertility is not being modelled well. It looks more like a smooth trend across field than a step function.



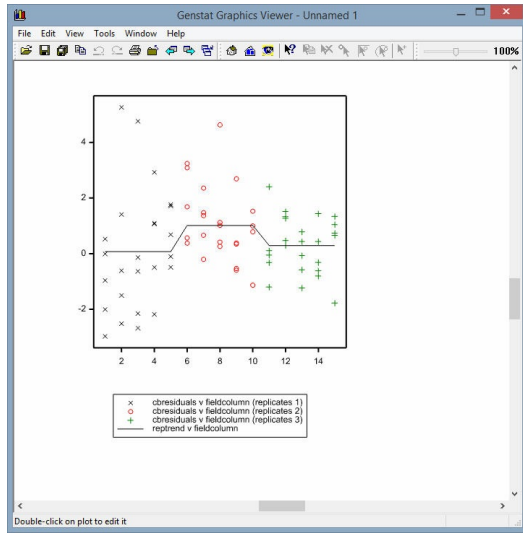


Figure 3.1

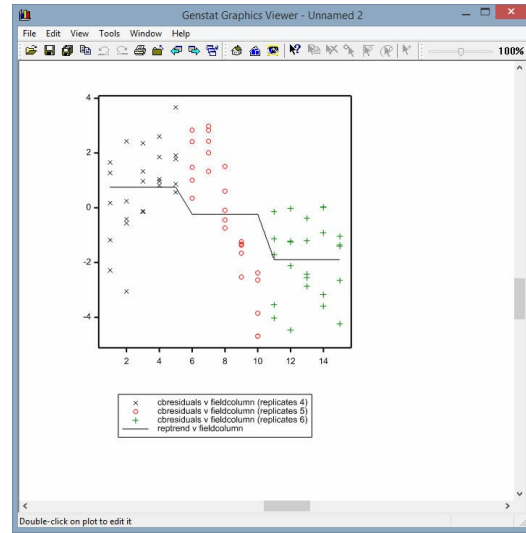


Figure 3.2

The conclusion is that randomized-block designs with large blocks may be insufficiently flexible to model fertility trends well. One solution is to use a more sophisticated design, like the lattice square actually used here. However, such designs are only available for specific numbers of treatments (in this case squares of integer i.e. 9, 16, 25, 36 and so on). The alternative, discussed in this chapter, is to use a more sophisticated analysis.

## 3.2 Correlation modelling

Traditional analysis of variance is based on “Fisher's 3 R's”, namely

- replication – it is usually recommended to replicate all treatments,
- randomization – on any pair of plots there should be an equal chance of getting any pair of treatments (this guarantees the validity of analysis, lack of bias etc.),
- blocking – group similar plots together, and fit a random term to model the differences between the groups (inducing a uniform correlation on the members of each group).

In contrast, in spatial analysis

- often only check (or control) treatments are replicated,
- you randomize where possible (but design may constrain which treatments appear on some of the plots), and
- you take account of variation by fitting models to describe how the correlation between each plot and its neighbours changes as the neighbours get further away.

As explained in Section 1.1, the traditional mixed model is

$$y = \sum_i \mathbf{X}_i \beta_i + \sum_i \mathbf{Z}_i u_i + \varepsilon$$

where

$y$  is the vector of data values,

$\beta_i$  is the vector of fixed effects for treatment term  $i$  with design matrix  $\mathbf{X}_i$ ,

$u_i$  is the vector of random effects for random term  $i$  with design matrix  $\mathbf{Z}_i$ ,

$\varepsilon$  is the vector of residuals.

Each element of the residual vector  $\varepsilon$  follows a Normal distribution with mean zero and variance  $\sigma^2$ . Equivalently, we can say that the vector  $\varepsilon$  follows a multivariate Normal distribution with a mean vector of zeros, and variance-covariance matrix  $\sigma^2 \mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix.

Likewise, each element of the vector of random effects  $u_i$  follows a Normal distribution with mean zero and variance  $\gamma_i \sigma^2$ . Again, equivalently, we can say that the vector  $u_i$  follows a multivariate Normal distribution with mean vector of zeros, and variance-covariance matrix  $\gamma_i \sigma^2 \mathbf{I}$ .

In correlation modelling, the equation of the mixed model remains the same, but the vectors of random effects  $u_i$  now follow multivariate Normal distributions with a variance-covariance matrix  $\gamma_i \sigma^2 \mathbf{G}$ , where the matrix  $\mathbf{G}$  is defined using a correlation model (or it may remain the identity matrix  $\mathbf{I}$  if the effects are independent, i.e. uncorrelated, as in traditional model). Likewise, the residual vector  $\varepsilon$  now follows a multivariate Normal distribution with variance  $\sigma^2 \mathbf{R}$ , where the matrix  $\mathbf{R}$  may be defined using a correlation model. (Again  $\mathbf{R}$  remains the identity matrix  $\mathbf{I}$  if the effects are independent.)

The full range of correlation models is defined in Section 3.3. If we write the value in the correlation matrix  $\mathbf{C}$  (either  $\mathbf{G}$  or  $\mathbf{R}$ )  $c_{ij}$  in row  $i$  and column  $j$  as  $c_{ij}$ , the most useful models for spatial modelling can be defined as follows:

identity	$c_{i,i} = 1$ $c_{i,j} = 0$ , for $i \neq j$
auto-regressive order 1 (AR1)	$c_{i,i} = 1$ $c_{i+k,i} = \phi^k$
auto-regressive order 2 (AR2)	$c_{i,i} = 1$ $c_{i+1,i} = \phi_1 / (1 - \phi_2)$ $c_{i,j} = \phi_1 c_{i-1,j} + \phi_2 c_{i-2,j}$ $i > j+1$ , $-1 < \phi_1, \phi_2 < 1$ , $ \phi_1 + \phi_2  < 1$ , $\phi_2 - \phi_1 < 1$ , $\phi_2 > -1$
power-distance	$c_{i,i} = 1$ $c_{i,j} = \phi^d$ $d =  i-j $

Notice that the AR1 and power-distance models are identical if the plots are equally spaced.

In a 2-dimensional spatial model, a correlation model is fitted to a random term `fieldrow.fieldcolumn`, where `fieldrow` and `fieldcolumn` are factors representing row and column positions up-and-down or from side-to-side of the whole field (rather than within replicates). Usually a *separable correlation model* is fitted, in which the correlation between the plots at coordinates  $(i,j)$  and  $(k,l)$  is the product of a correlation from a model defined on the rows of the experiment, and a correlation from a model defined on the columns of the experiment: i.e.

correlation  $cr_{ik}$  between rows  $(i-k)$  apart

× correlation  $cc_{jl}$  between columns  $(j-l)$  apart

with the correlations  $cr_{ik}$  and  $cc_{jl}$  being defined by one of the models above. Separable correlation models are often represented using the *direct product* symbol  $\otimes$ : so a model constructed from two AR1 models is written as  $\text{AR1} \otimes \text{AR1}$ .

The models can all be defined using the `VSTRUCTURE` directive, as explained in Section 3.3, but this is easier using the `REML` menus. To open the menu to use when the plots are on a regular grid, you click **Stats** on the menu bar, select **Mixed Models (REML)**, select **Spatial Models** and then click on **Regular Grid** (see Figure 3.3), to produce the menu in Figure 3.4.

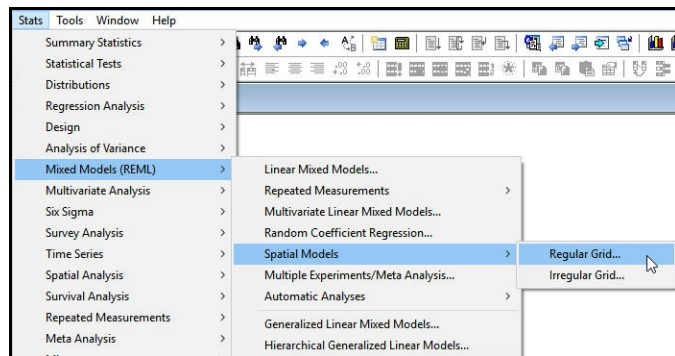


Figure 3.3

We set the Data variate to `yield` and the Fixed terms to `variety`, and then consider the error model. We now need to specify the row and column factors (in the **Row factor** and **Column factor** boxes), and select the required correlation model from those available in the **Row-model** and **Column-model** boxes. Here we set the **Row-model** and **Column-model** list boxes to **AR order 1**. Other boxes allow you to

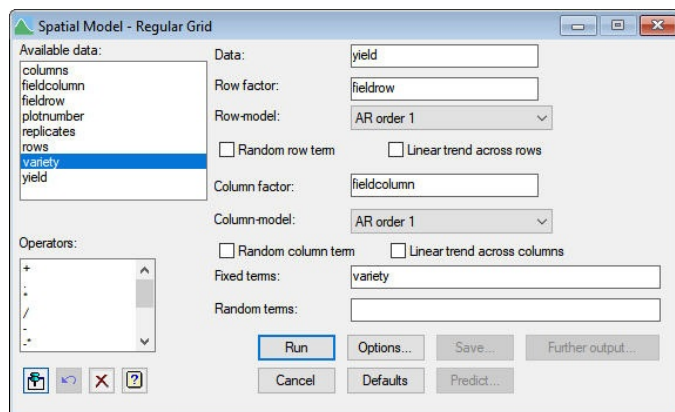


Figure 3.4

fit a random row effect in addition to the row model, or to fit a linear trend across rows, and there are similar boxes for the columns. You can also specify additional random terms, in the **Random terms** box. These might include other types of blocking, as for example if the plots had been sown or harvested on different days or by different operators.

Clicking on **Run** produces the output below. The default output contains sections to describe the covariance structures (i.e. models) that are being fitted, and give their parameter estimates. The fact that rather different estimates are produced for the autoregressive parameter in the two directions arises from the fact that the plots of the design were not square.

## REML variance components analysis

Response variate:	yield
Fixed model:	Constant + variety
Random model:	fieldrow.fieldcolumn
Number of units:	150

fieldrow.fieldcolumn used as residual term with covariance structure as below

Sparse algorithm with AI optimisation

## Covariance structures defined for random model

Covariance structures defined within terms:

Term	Factor	Model	Order	No. rows
fieldrow.fieldcolumn	fieldrow	Auto-regressive (+ scalar)	1	10
	fieldcolumn	Auto-regressive	1	15

## Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
fieldrow.fieldcolumn			Sigma2	3.876	0.775
	fieldrow	AR(1)	phi_1	0.4586	0.0826
	fieldcolumn	AR(1)	phi_1	0.6838	0.0633

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	313.04	24	13.04	80.0	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	313.04	24	13.04	80.0	<0.001

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

---

The fact that the F statistic in the spatial analysis is larger than the F statistic from the analysis of variance of the lattice square design suggests that the spatial analysis is modelling the fertility of the field more effectively. However, as we shall see later in this section, it is wrong to select the random model by seeing how significant we can make the treatments! In fact you should establish the appropriate random model before you start to consider the fixed model.

You can assess the effect of extending or simplifying the random model by seeing how this changes the deviance. Criteria like the Akaike and Schwarz Bayes information coefficients can be also used, and this can be done even for models where one is not an extension of the other one. These coefficients each involve the deviance plus a term that takes account of the numbers of parameters in each of the models (in ways that differ between the two coefficients). For the random model, it is best to leave the box [Use full likelihood for AIC/SIC](#) unchecked, and use the [REML](#) deviance. You should use the full likelihood if you want to assess changes in the fixed model (while keeping the random model unchanged). Details are given in the on-line help for the [VAIC](#) procedure, or its description in Part 3 of the *Genstat Reference Manual*.

In Figure 3.5 we use the [Spatial Model Further Output](#) menu to display the deviance and the two coefficients. The notes printed in the output emphasize that you can compare the random models only if the fixed model is unchanged.

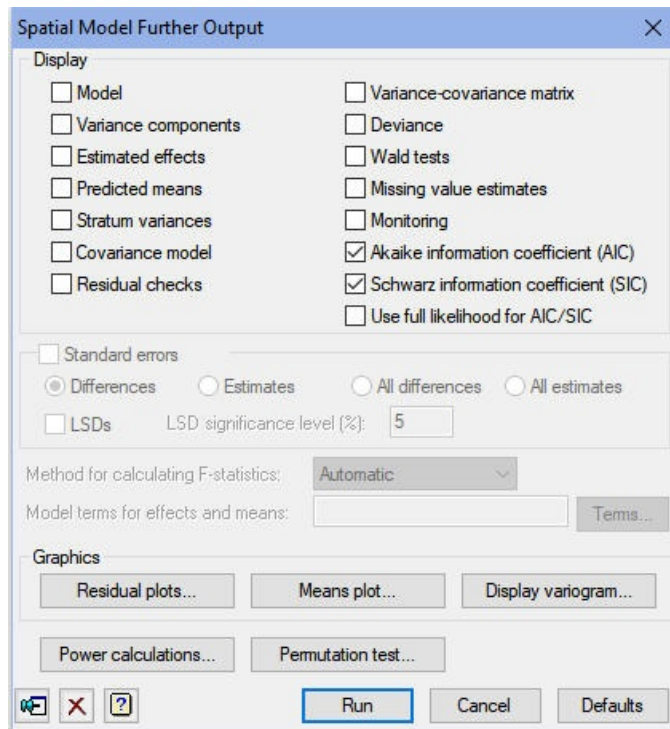


Figure 3.5

## Deviance: $-2 \times \text{Log-Likelihood}$

Deviance	d.f.
249.35	122

Note: deviance omits constants which depend on fixed model fitted.

Akaike information coefficient	255.35
Schwarz Bayes information coefficient	263.84

Note: omits constants,  $(n-p)\log(2\pi) - \log(\det(X'X))$ , that depend only on the fixed model.

(based on the residual log-likelihood)

We now assess whether we need a more complicated correlation structure, with AR order 2 models on the rows and columns. To simplify the output, we first click on **Options** to bring up the **Spatial Model Options** menu (Figure 3.6), modify the **Display** settings so that only the **Deviance**, **Akaike information coefficient (AIC)** and **Schwarz information coefficient (SIC)** boxes are checked, and click on **OK**. In the main menu (Figure 3.4), we then change the **Row-model** and **Column-model** list boxes to **AR order 2**, and click on **Run**.

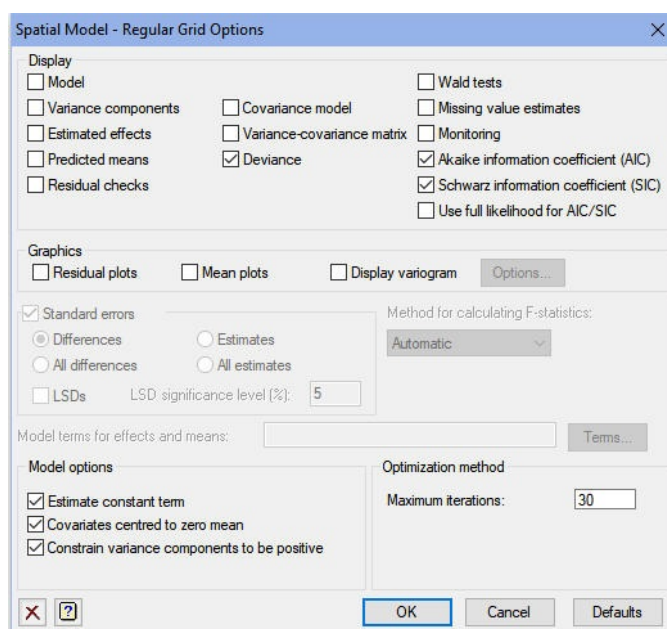


Figure 3.6

## Deviance: -2\*Log-Likelihood

Deviance	d.f.
246.35	120

Note: deviance omits constants which depend on fixed model fitted.

Akaike information coefficient	256.35
Schwarz Bayes information coefficient	270.49

Note: omits constants,  $(n-p)\log(2\pi) - \log(\det(X'X))$ , that depend only on the fixed model.

(based on the residual log-likelihood)

Changes in the deviance are approximately distributed as  $\chi^2$ , so here the extra variance parameters for the order 2 models have a  $\chi^2$  value of  $249.35 - 246.35 = 3.00$  on  $122 - 120 = 2$  degrees of freedom. There is thus no evidence that we need auto-regressive structures of order 2 rather than 1. Note, however, that the absolute value of the deviance value is not useable. To simplify the calculations some constant terms (which depend only on the fixed model) are omitted, so the value printed by Genstat may even be negative.

The Akaike and Schwarz Bayes information coefficients lead to the same conclusion: the best model is the one with the smaller coefficient and, for both of these, that is model with AR order 1 structures. In most situations, the conclusions from the two coefficients will agree. The difference is that the Schwarz Bayes information coefficient tends to weight more heavily against the inclusion of extra parameters in the model.



Next we look to see if there is any need to include a term for measurement error. We can do this by specifying a random term that indexes the individual units of the design. In `Slatehall.gsh` we have a factor `plotnumber` which has values 1 - 150 running over the 150 plots in the experiment. Note that we cannot use the term `fieldrow.fieldcolumn`, as this is already being as the spatial covariance term. If you do not

have a factor like `plotnumber`, the `REML` directive allows you to use the string `'*units*'` to denote an internal factor with a level for every units of the design. We enter `plotnumber` into the **Random terms** box (see Figure 3.7), and again click on **Run**.

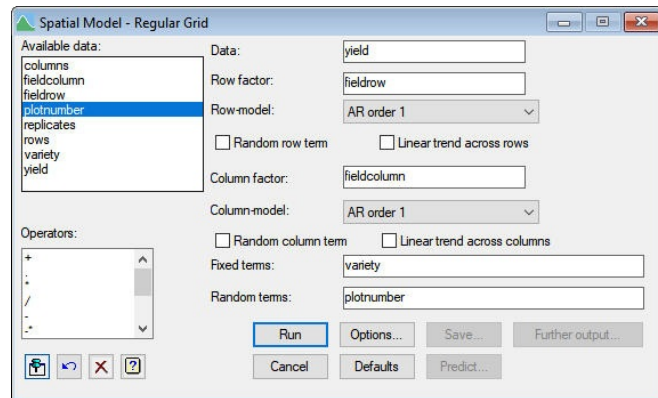


Figure 3.7

*Warning 1, code VC 53, statement 1 on line 191*

Command: REML [PRINT=deviance; MAXCYCLE=20; MVINCLUDE=explanatory,yvariate; METH

More than one residual term specified - first term found will be used as R.

### Deviance: -2\*Log-Likelihood

Deviance	d.f.
242.35	121

Note: deviance omits constants which depend on fixed model fitted.

Akaike information coefficient	250.35
Schwarz Bayes information coefficient	261.67

Note: omits constants,  $(n-p)\log(2\pi) - \log(\det(X'X))$ , that depend only on the fixed model.

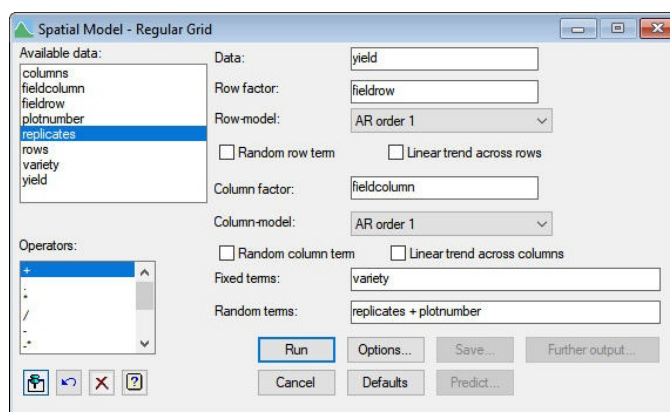
(based on the residual log-likelihood)

Genstat prints a warning to say that there are two potential-residual terms (i.e. terms corresponding to the matrix **R** defined earlier in this section). This can be ignored as here they are being used for different purposes: `plotnumber` represents the residual (plot) variation, and `fieldrow.fieldcolumn`, is a spatial co term.

The difference in deviances is  $249.35 - 242.35 = 7$  on 1 degree of freedom. So it appears that we do need to include measurement error in the analysis.



The **Random terms** box also allows you to include other types of blocking. For example, the **replicates** term might be needed if it represented differences in fertility beyond those described by the spatial model or if represented differences in the husbandry of the plots, as e.g. if different replicates had been sown or harvested on different days or by different operators.



**Figure 3.8**

You should also include random terms if they were involved in the allocation and randomization of a fixed term. For example, in the split-plot analysis in Section 1.1, the **blocks.wplots** term should be included, as the varieties were applied and randomized on the whole-plots within each block, rather than to the individual sub-plots of the design. If **blocks.wplots** is not included in the random model, the wrong denominator degrees of freedom will be used in the tests for fixed effects. (A case could be made for including all the conventional random terms here too, as the allocation of the varieties did depend on blocking structure of the balanced-lattice design. However, comparing the analyses here with the one in Section 1.8 shows that this would not make any substantial difference to the denominator degrees of freedom. A clue as to why this is true is given by the fact that most of the information on the varieties is in the bottom stratum of the analysis of variance.)

The **replicates** term is added to the random terms in Figure 3.8, and the output below shows that the deviance is unchanged.

### *Warning 2, code VC 53, statement 1 on line 207*

Command: REML [PRINT=deviance; MAXCYCLE=20; MVINCLUDE=explanatory,yvariate;  
METH  
More than one residual term specified - first term found will be used as R.

### Deviance: -2\*Log-Likelihood

Deviance	d.f.
242.35	120

Note: deviance omits constants which depend on fixed model fitted.

Akaike information coefficient	252.35
Schwarz Bayes information coefficient	266.49

Note: omits constants,  $(n-p)\log(2\pi) - \log(\det(X'X))$ , that depend only on the fixed model.

(based on the residual log-likelihood)

This might seem a little surprising, but the reason becomes clear if we use the [Spatial Model Further Output](#) menu (Figure 3.9) to display the variance components. By default, the `REML` menus constrain the variance components to be positive, and here the component for `replicates` is stuck on the boundary.

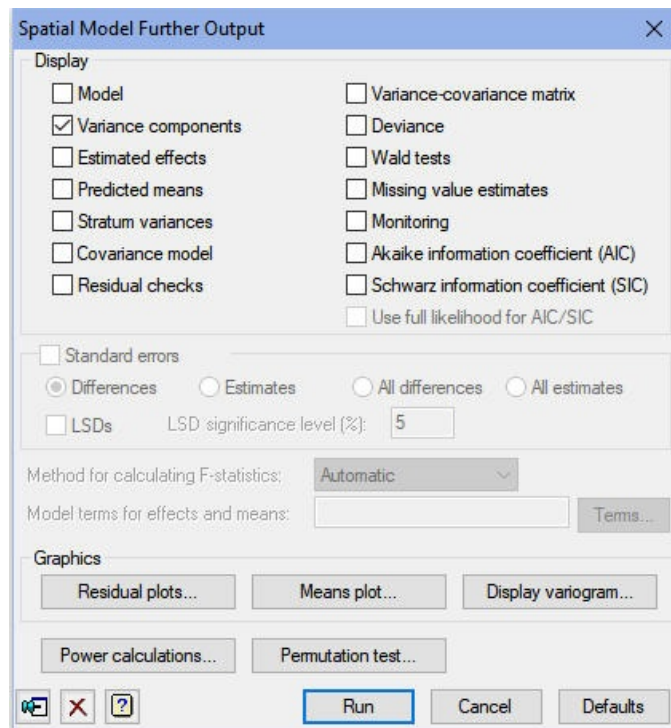


Figure 3.9

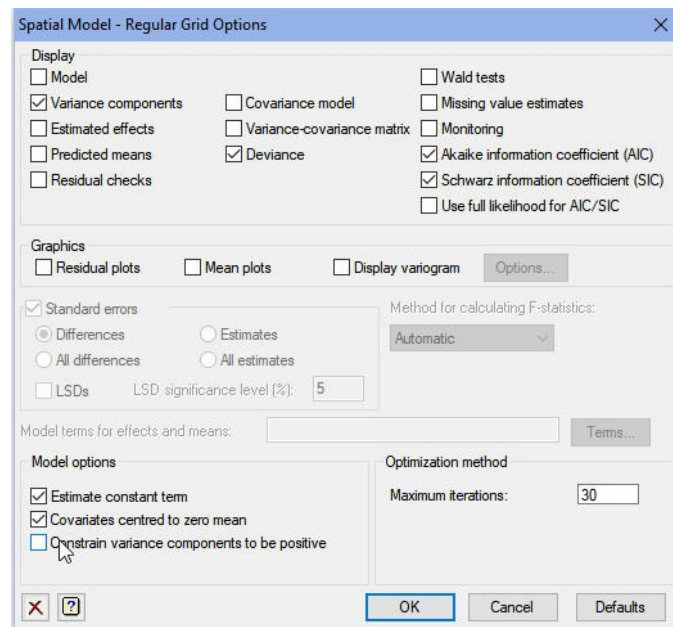
### Estimated variance components

Random term	component	s.e.
replicates	0.000	bound
plotnumber	0.486	0.179

### Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
fieldrow.fieldcolumn			Sigma2	4.580	1.670
	fieldrow	AR(1)	phi_1	0.6827	0.1023
	fieldcolumn	AR(1)	phi_1	0.8438	0.0684

We can remove the constraint by unchecking the relevant box in the **Spatial Model Options** menu (Figure 3.10) and rerun the analysis printing the deviance, the coefficients and the variance components.



**Figure 3.10**

*Warning 3, code VC 53, statement 1 on line 224*

Command: REML [PRINT=components,deviance; MAXCYCLE=20;  
MVINCLUDE=explanatory,yva  
More than one residual term specified - first term found will be used as R.

## Estimated variance components

Random term	component	s.e.
replicates	-0.078	0.162
plotnumber	0.500	0.181

## Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
fieldrow.fieldcolumn			Sigma2	4.662	1.707
	fieldrow	AR(1)	phi_1	0.6922	0.1022
	fieldcolumn	AR(1)	phi_1	0.8442	0.0685

## Deviance: -2\*Log-Likelihood

Deviance	d.f.
242.20	120

Note: deviance omits constants which depend on fixed model fitted.

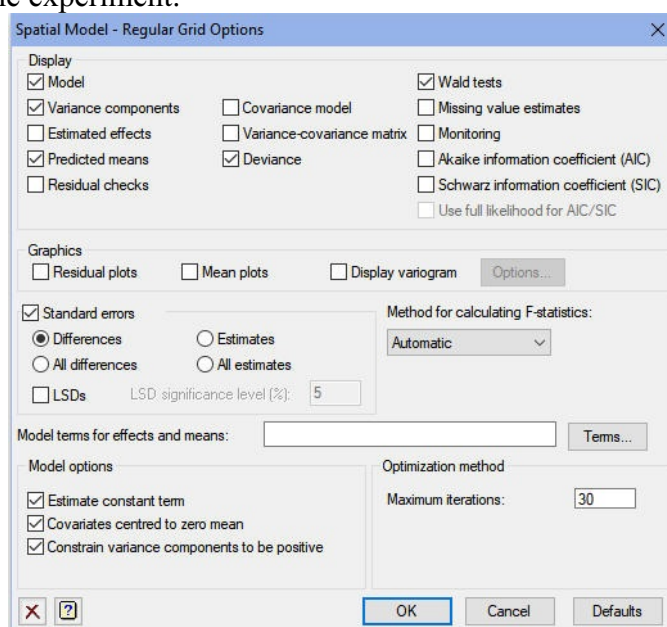
Akaike information coefficient	252.20
Schwarz Bayes information coefficient	266.34

Note: omits constants,  $(n-p)\log(2\pi) - \log(\det(X'X))$ , that depend only on the fixed model.

(based on the residual log-likelihood)

The variance component for `replicates` is negligible. So the replicate differences in this experiment arise only from fertility trends, and these have been described by spatial correlation structure. (Again the Akaike and Schwarz Bayes information coefficients lead to the same conclusion.) However, we should still include `replicates` as this was a factor involved in the design of the experiment.

We can now move on to assess the terms in the fixed model. So we set the **Spatial Model Options** menu to print the model, variance components, predicted means and Wald tests (Figure 3.11), generating the output shown below.



**Figure 3.11**

*Warning 4, code VC 53, statement 1 on line 240*

Command: REML [PRINT=model,components,means,waldTests; MAXCYCLE=20;  
FMETHOD=auto

More than one residual term specified - first term found will be used as R.

## REML variance components analysis

Response variate: yield  
Fixed model: Constant + variety  
Random model: replicates + fieldrow.fieldcolumn + units  
Number of units: 150

fieldrow.fieldcolumn used as residual term with covariance structure as below

Sparse algorithm with AI optimisation

## Covariance structures defined for random model

Covariance structures defined within terms:

Term	Factor	Model	Order	No. rows
fieldrow.fieldcolumn	fieldrow	Auto-regressive (+ scalar)	1	10
	fieldcolumn	Auto-regressive	1	15

## Estimated variance components

Random term	component	s.e.
replicates	-0.078	0.162
units	0.500	0.181

## Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
fieldrow.fieldcolumn			Sigma2	4.622	1.707
	fieldrow	AR(1)	phi_1	0.6922	0.1022
	fieldcolumn	AR(1)	phi_1	0.8442	0.0685

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	241.61	24	10.05	74.7	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	241.61	24	10.05	74.7	<0.001

## Table of predicted means for Constant

14.47 Standard error: 0.894

## Table of predicted means for variety

variety	1	2	3	4	5	6	7	8
	12.45	15.14	14.05	14.06	14.71	15.21	13.74	14.53
variety	9	10	11	12	13	14	15	16
	12.63	11.95	13.29	14.40	16.24	13.00	14.69	12.86
variety	17	18	19	20	21	22	23	24
	14.94	15.28	16.50	16.44	15.18	16.10	13.18	15.55

variety            25  
                     15.73

Standard errors of differences

Average:                    0.6072  
Maximum:                   0.6401  
Minimum:                   0.5719

Average variance of differences: 0.3689

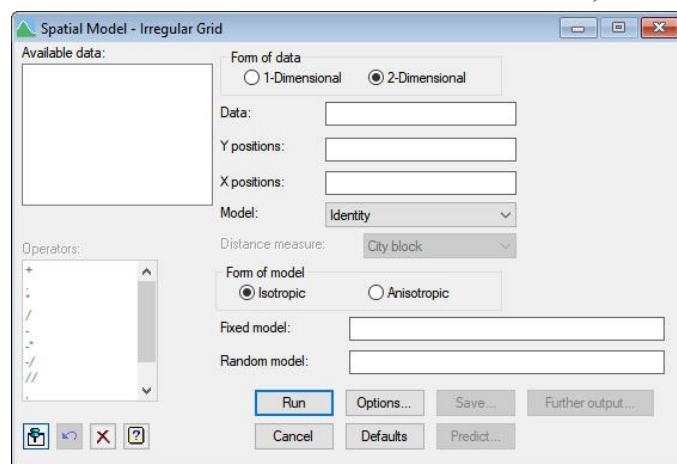
The output describes the models, and prints their parameters. Notice that the Wald and F statistics (now 241.61 and 10.05) are smaller than those (313.04 and 13.04) from the original analysis in this section that fitted just the AR order 1  $\times$  AR order 1 spatial model. The inclusion of measurement error is needed to represent the fertility patterns in the field correctly, but this acknowledges that the data are more variable than can be modelled just by the spatial model. This emphasizes the point that you should not select your random model by seeing which one gives the most significant results for the fixed terms!

So, although it is interesting to notice the increase in the F statistic (from 8.84 to 10.21) compared to the previous, conventional analysis of the data as a lattice square design (Section 1.8). The correct reason for using the spatial analysis is that, if we had used the similar [Linear Mixed Models Further Output](#) menu at in Section 1.8 to display the Akaike and Schwarz Bayes coefficients we would have obtained the values

Akaike information coefficient	272.28
Schwarz Bayes information coefficient	283.59

and these are much larger than the values 252.20 and 266.34 above. (Remember that we cannot use the deviance here, as one model is not an extension of the other one.)

The menu for an irregular grid (Figure 3.12) is similar to the menu for a regular grid (Figure 3.4), but with irrelevant boxes removed and the boxes for row and column factors replaced by boxes for x and y positions (or coordinates). The model choices are restricted to identity or power. The power can be of "city block" distance (row distance plus column distance), or squared distances (row distance squared plus column distance squared) or



**Figure 3.12**

Euclidean distance (square root of row distance squared plus column distance squared).

### 3.3 The VSTRUCTURE directive

Correlation models are defined by the `VSTRUCTURE` directive. The spatial analysis menus use only a limited selection of its facilities, so we present the full syntax below. However, for a full description you should read the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 5.4.

---

#### VSTRUCTURE directive

Defines a variance structure for random effects in a `REML` model.

##### Options

<code>TERMS</code> = <i>formula</i>	Model terms for which the covariance structure is to be defined
<code>FORMATION</code> = <i>string token</i>	Whether the structure is formed by direct product construction or by definition of the whole matrix ( <code>direct</code> , <code>whole</code> ); default <code>dire</code>
<code>CORRELATE</code> = <i>string token</i>	Whether to impose correlation across the model terms if several are specified ( <code>none</code> , <code>positivedefinite</code> , <code>unrestricted</code> ); default <code>none</code>
<code>CINITIAL</code> = <i>scalars</i>	Initial values for covariance matrix across terms
<code>COORDINATES</code> = <i>matrix or variates</i>	Coordinates of the data points to be used in calculating distance-based models

##### Parameters

<code>MODELTYPE</code> = <i>string tokens</i>	Type of covariance model associated with the term(s), or with individual factors in the term(s) if <code>FORMATION=direct</code> ( <code>identity</code> , <code>fixed</code> , <code>AR</code> , <code>MA</code> , <code>ARMA</code> , <code>power</code> , <code>boundedlinear</code> , <code>circular</code> , <code>spherical</code> , <code>linearvariance</code> , <code>banded</code> , <code>correlation</code> , <code>antedependence</code> , <code>unstructured</code> , <code>diagonal</code> , <code>uniform</code> , <code>FA</code> , <code>FAequal</code> ) default <code>iden</code>
<code>ORDER</code> = <i>scalar</i>	Order of model
<code>HETEROGENEITY</code> = <i>string token</i>	Heterogeneity for correlation matrices ( <code>none</code> , <code>outside</code> ); default <code>none</code>
<code>METRIC</code> = <i>string token</i>	How to calculate distances when <code>MODELTYPE=power</code> ( <code>cityblock</code> , <code>squared</code> , <code>euclidean</code> ); default <code>city</code>
<code>FACTOR</code> = <i>factors</i>	Factors over which to form direct products
<code>MATRIX</code> = <i>identifiers</i>	To define matrix values for a term or the factors when <code>MODELTYPE=fixed</code>
<code>INVERSE</code> = <i>identifiers</i>	To define values for matrix inverses (instead of the fixed matrices themselves) when <code>MODELTYPE=fixed</code>
<code>DISTANCES</code> = <i>symmetric matrices</i>	Symmetric matrix of pre-formed distances to be



	used in distance-based models of order one
<code>COORDINATES</code> = <i>matrices, variates or pointers</i>	Specifies coordinates of each factor level to be used in calculating distance-based models
<code>INITIAL</code> = <i>scalars, variates, matrices, symmetric matrices or pointers</i>	Initial parameter values for each correlation matrix (supplied in the structures appropriate for the model concerned)
<code>CONSTRAINTS</code> = <i>texts</i>	Texts containing strings <code>none</code> , <code>fix</code> or <code>positive</code> to define constraints for the parameters in each model
<code>EQUALITYCONSTRAINTS</code> = <i>variates</i>	Non-zero values in the variate indicate groups of parameters whose values are to be constrained to be equal

---

`VSTRUCTURE` can be used only after `VCOMPONENTS` has defined the fixed and random models. It can be used more than once to define different structures for different random terms. The information is accumulated within Genstat, and it will all be used by subsequent `REML` commands. You can check on the model and covariance structures defined at any time by using the `VSTATUS` directive. To cancel a covariance structure for a term you simply need to use `VSTRUCTURE` to change the model back to the default identity matrix. To cancel all covariance structures you can give a new `VCOMPONENTS` command and redefine the fixed and random models.

By default the covariance model is formed as the direct product of several matrices, one for each factor in the term. The models used to define the matrices are specified by the first parameter of `VSTRUCTURE`, called `MODELTYPE`, and the `FACTOR` parameter specifies the corresponding factor. Setting option `FORMATION=whole` allows you to define a single covariance matrix for the whole term. (You would then omit the `FACTOR` parameter.)

Other parameters specify information for the model, for example `ORDER` specifies its order. The `TERM` option specifies the model term. So, in particular, the `AR1⊗AR1` structure, used for `fieldrow.fieldcolumn` in Section 1.2, would be defined by

```
VSTRUCTURE [fieldrow.fieldcolumn] AR,AR;\
FACTOR=fieldrow,fieldcolumn; ORDER=1,1
```

and the `AR2⊗AR2` structure would be defined by

```
VSTRUCTURE [fieldrow.fieldcolumn] AR,AR;\
FACTOR=fieldrow,fieldcolumn; ORDER=2,2
```

Any factors in the term for which no model is defined are assumed to have an identity matrix as their correlation matrix (i.e. their effects are uncorrelated). This actually means that you can specify models for several terms at once, but it is more usual to specify only one.

If we write the value in the correlation matrix **C** in row *i* and column *j* as  $c_{ij}$ , the matrices corresponding to the full range of settings of `MODELTYPE` can be defined as follows.

identity	identity matrix	$c_{i,i} = 1, c_{i,j} = 0, \text{ for } i \neq j$
fixed	fixed matrix	$c_{i,j}$ specified
AR	auto-regressive order 1 or 2 ( $\varphi_2=0$ for order 1)	$c_{i,i} = 1$ $c_{i+1,i} = \varphi_1 / (1-\varphi_2)$ $c_{i,j} = \varphi_1 c_{i-1,j} + \varphi_2 c_{i-2,j},$ $i > j+1$ $-1 < \varphi_1$ $\varphi_2 < 1$ $ \varphi_1 + \varphi_2  < 1$ $\varphi_2 - \varphi_1 < 1$ $\varphi_2 > -1$
MA	moving average order 1 or 2 ( $\theta_2=0$ for order 1)	$c_{i,i} = 1$ $c_{i+1,i} = -\theta_1(1-\theta_2)/(1+\theta_1^2+\theta_2^2)$ $c_{i+2,i} = -\theta_2 / (1+\theta_1^2+\theta_2^2)$ $c_{i,j} = 0, i > j+2$ $-1 < \theta_1$ $\theta_2 < 1$ $\theta_2 \pm \theta_1 < 1$
ARMA	auto-regressive moving-average order 1	$c_{i,i} = 1$ $c_{i+1,i} = (\theta - \varphi)(1 - \varphi\theta) /$ $(1 + \theta^2 - 2\varphi\theta)$ $c_{i,j} = \varphi c_{i-1,j}, i > j+1$ $-1 < \varphi,$ $\theta < 1$
power	based on distance order 1 or 2 ( $\varphi_1 = \varphi_2$ for order 1)	$c_{i,i} = 1$ $c_{i,j} = \varphi_1^{d_1} \varphi_2^{d_2}$ $d_1, d_2 = \text{distance in 1st and}$ $2\text{nd dimensions}$ $0 < \varphi_1$ $\varphi_2 < 1$
boundedlinear	based on distance order 1	$c_{i,j} = 1 - d/\varphi \text{ for } d \leq \varphi,$ $c_{i,j} = 0 \text{ for } d > \varphi$ $0 < \varphi$
circular	based on distance order 1	$c_{i,j} = 1 -$ $(2/\pi) \{ (d/\varphi) \sqrt{1 - (d/\varphi)^2} +$ $\sin^{-1}(d/\varphi) \} \text{ for } d \leq \varphi,$ $c_{i,j} = 0 \text{ for } d > \varphi$ $0 < \varphi$

spherical	based on distance order 1	$c_{i,j} = 1 - 1.5(d/\phi) + 0.5(d/\phi)^3$ for $d \leq \phi$ , $c_{i,j} = 0$ for $d > \phi$ $0 < \phi$
linearvariance	based on distance order 1	$c_{i,i} = 1 - 2\phi d / \max(d)$ $0 < \phi < 1$
banded	equal bands $1 < \text{order} < \text{nrows} - 1$	$c_{i,i} = 1$ $c_{i+k,i} = \theta_k, 1 < k < \text{order}$ $c_{i+k,i} = 0$ , otherwise $-1 < \theta_k < 1$
correlation	general correlation matrix $1 < \text{order} < \text{nrows} - 1$	$c_{i,i} = 1$ $c_{i,j} = \theta_{ij}, 1 <  i-j  \leq \text{order}$ $c_{i,j} = 0,  i-j  > \text{order}$ $-1 < \theta_{ij} < 1$
uniform	uniform matrix	$c_{i,j} = \theta$ for all $i,j$
diagonal	diagonal matrix	$c_{i,i} = \theta_i$ $c_{i,j} = 0, i \neq j$
antedependence	ante-dependence model $1 < \text{order} < \text{nrows} - 1$	$\mathbf{C}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}'$ $\{\mathbf{D}\}_{i,i}^{-1} = d_i^{-1}$ , $\{\mathbf{D}\}_{i,j} = 0$ for $i \neq j$ $\{\mathbf{U}\}_{i,i} = 1$ , $\{\mathbf{U}\}_{i,j} = u_{ij}$ , $1 \leq j-i \leq \text{order}$ $\{\mathbf{U}\}_{i,j} = 0$ , for $i > j$
unstructured	general covariance matrix $1 < \text{order} < \text{nrows} - 1$	$c_{i,j} = \theta_{ij}, 0 <  i-j  \leq \text{order}$ $c_{i,j} = 0,  i-j  > \text{order}$
FA	factor analytic order = 1 or 2	$\mathbf{C} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ $\mathbf{\Lambda}$ is an $\text{nrows} \times q$ matrix order= $q$ $\{\mathbf{\Psi}\}_i = \psi_i$ for $i=1 \dots \text{nrows}$
FAequal	factor analytic with common variance order = 1 or 2	$\mathbf{C} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ $\mathbf{\Lambda}$ is an $\text{nrows} \times q$ matrix order= $q$ $\{\mathbf{\Psi}\}_i = \psi$ for $i=1 \dots \text{nrows}$

Initial parameter values can be specified using the `INITIAL` parameter. For most models, the number of initial values required is the number of parameters, and default values will be generated. However, for `unstructured` models, a full covariance matrix of initial values must be given, and for the `correlation` model a full correlation matrix

must be provided. For the `ante-dependence` model, either a full covariance matrix can be provided, or a pointer to a  $\mathbf{U}$  and a  $\mathbf{D}^{-1}$  matrix of the correct forms. For the `FA` and `FAequal` models, a pointer must be used to give the initial  $\mathbf{\Lambda}$  and  $\mathbf{\Psi}$  matrices, otherwise default initial values are generated. The `FAequal` model can be used to get initial values for the `FA` model. Initial values are required for these models because the algorithm may not converge when many parameters are fitted if the starting values are not realistic. Initial values might be generated from covariance matrices estimated by fitting simpler models, or from residuals from a null variance model. A missing value in the initial values is taken to mean that the value is inestimable and it will be fixed at a small value for the analysis. Alternatively, a parameter can be fixed at its initial value using the `CONSTRAINTS` parameter. The codes (not case sensitive and able to be abbreviated) may take value `fix` to indicate the parameter is to be fixed at its initial value, `positive` to indicate it is to remain positive or `none` to indicate no constraints. The default is positive/no constraint depending on context, for example scaling parameters are always constrained to remain positive. The default is positive/no constraint depending on context, for example scaling parameters are always constrained to remain positive. The `EQUALITYCONSTRAINTS` parameter allows you to constrain some of the parameters to have the same value. The variate that it specifies contains a zero value if there is no constraint, and an identical integer value for any set of parameters whose values are to be equal. So, a variate containing the values (0,1,2,1,2) would constrain the second parameter to be equal to the fourth parameter, and the third parameter to be equal to the fifth parameter.

It may sometimes be desirable to allow for unequal variances for the models defined in terms of correlation matrices: that is, for the `AR`, `MA`, `ARMA`, `uniform`, `power`, `boundedlinear`, `circular`, `spherical`, `linearvariance`, `banded` and `correlation` models. This can be done by setting option `HETEROGENEITY=outside`. This means a diagonal matrix  $\mathbf{D}$  of standard errors will be applied to the correlation matrix  $\mathbf{C}$  to generate a matrix  $\mathbf{D}^{1/2}\mathbf{C}\mathbf{D}^{1/2}$ . In this case, a number of extra parameters (equal to the number of effects in the factor or term) should be added to the vector of initial values. These models allow investigation of a structured correlation pattern for changing variances and are particularly useful in the analysis of repeated measurements data when variance increases over time. For example, to allow for changing variance over time in our example above, we can specify

```
VCOMPONENTS [FIXED=Tmt] RANDOM=Subject.Week
VSTRUCTURE [TERM=Subject.Week] MODELTYPE=AR; ORDER=1;\
    FACTOR=Week; HETEROGENEITY=outside
REML Y
```

In some circumstances, you may wish to define a single model to apply to the whole term, instead of using the direct product form illustrated above. In this case, you should set option `FORM=whole`. Note that, when a term consists of a single factor, it is not necessary to set the `FACTOR` option.

With `MODELTYPE=fixed`, you must either use the `MATRIX` option to specify the values of the covariance matrix  $\mathbf{C}$ , or the `INVERSE` option to specify the inverse matrix. Values for the matrix or its inverse can be supplied as diagonal matrices or symmetric matrices. In addition, values for the inverse matrix can be supplied in sparse form as a pointer. The output from `VPEDIGREE` is designed for input here, but you can also define the inverse matrix explicitly. The second element of the pointer should then be a variate containing

the non-zero values of the inverse in lower triangular order. The first element should be a factor, with number of levels equal to the number of rows  $n(n+1)/2$  of the matrix. This has firstly a block of  $n$  values giving the position in the variate of the first value stored for each row. There is then a block of values for each row in turn, giving the columns in which each non-zero value appears.

When `MODELTYPE=power` is used to define a distance-based model, the model can be of order 1 (isotropic) or 2 (anisotropic). For models with `ORDER=1`, a single set of distances must be formed. The necessary information can be supplied using either the `COORDINATES` option, or the `COORDINATES` parameter, or the `DISTANCES` parameter. With the `COORDINATES` option you can specify either a matrix, or a list of variates, to define multi-dimensional coordinates for each unit of the data. The length of the variates, or the number of rows of the matrix, must be equal to the number of data values. The number of variates, or the number of columns of the matrix, is equal to the number of dimensions. The coordinates for the levels of each `FACTOR` are then calculated as the mean values of the coordinates of the units included in the analysis with those levels. Alternatively, you can use the `COORDINATES` parameter to specify a single variate, a pointer to several variates or a matrix to define multi-dimensional coordinates for each level of the `FACTOR`. This parameter takes precedence over the `COORDINATES` option. The length of the variates, or the number of rows of the matrix, must be equal to the number of levels of the `FACTOR`. The number of variates, or the number of columns of the matrix, is again equal to the number of dimensions.

The distance calculation is defined by the `METRIC` option. For levels  $i$  and  $j$  with  $n$ -dimensional coordinates  $\{c_{ik}: k=1\dots n\}$  and  $\{c_{jk}: k=1\dots n\}$  the distance  $d_{ij}$  is defined as

$$\begin{aligned} d_{ij} &= \sum_k |c_{ik} - c_{jk}| && \text{for METRIC=cityblock (the default);} \\ d_{ij} &= \sum_k (c_{ik} - c_{jk})^2 && \text{for METRIC=squared; and} \\ d_{ij} &= \{\sum_k (c_{ik} - c_{jk})^2\}^{1/2} && \text{for METRIC=euclidean.} \end{aligned}$$

Finally, you can supply a symmetric matrix of pre-calculated distances, using the `DISTANCES` parameter, and this takes precedence over the `COORDINATES` parameter and option. The number of rows of the `DISTANCES` matrix must be equal to the number of levels of the `FACTOR`.

When `MODELTYPE=power` and `ORDER=2`, the `DISTANCES` parameter cannot be used, and only two-dimensional coordinates are allowed. The coordinates must be specified using either the `COORDINATES` option or parameter, as described above. The distances are calculated within each dimension separately, according to the setting of the `METRIC` option. In this case the Euclidean and city-block distances are equivalent.

The spherical family of geostatistical models correspond to the `MODELTYPE` settings `boundedlinear` (for one-dimensional distances), `circular` (for one or two dimensions) and `spherical` (for one or two dimensions). These models are based on distances, and require coordinates to be supplied using either the `COORDINATES` option (to give coordinates for each data value), or the `COORDINATES` parameter (to give coordinates for each factor level), as described for `MODELTYPE=power` above. The parameter  $\phi$  is interpreted as the range at which the correlation is considered to have decayed to zero. A small value therefore indicates weak correlation, and a large value indicates stronger correlation. These models do not have continuous second derivatives, and their log-likelihood may be multi-modal. To detect this potential problem, it is therefore important to start their estimation from several different initial values; this can be done using the `INITIAL` parameter as described above. To ensure that the estimated

correlation matrix differs from the identity matrix, it is necessary for the range parameter to be larger than the minimum distance specified by the coordinates; any initial value smaller than this will be adjusted.

The setting `MODELTYPE=linearvariance` specifies the linear variance model. This is parameterized so that the parameter  $\phi$  lies in the range  $[0,1]$ , which allows correlations in the range  $[-1,1]$ . Values of  $\phi$  close to one indicate weak correlation and values close to zero indicate strong correlation between neighbouring observations.

The `CORRELATE` option allows you to specify correlations between model terms which have equal numbers of effects. A common correlation will then be fitted between parallel effects. For example, consider a random coefficient regression model where the fixed model contains common response to covariate `x` and the random model allows for deviations in the intercept and slope about this line for each subject. The random intercept and slope for each subject may be correlated, but subjects are independent. This correlation across terms is defined using the `CORRELATE` option as follows:

```
VCOMPONENTS [FIXED=X] RANDOM=SUBJECT+SUBJECT.X
VSTRUCTURE [SUBJECT+SUBJECT.X; CORRELATE=positivedefinite;\
CINITIAL=(1,0.1,0.3); FORM=whole]
```

The `CORRELATE` option setting `positivedefinite` is used to ensure that the correlation matrix between the terms remains positive definite. This constraint can be relaxed using the setting `unrestricted` (an unstructured covariance matrix is then used to describe covariance across the terms). The model fitting is done here in terms of a covariance matrix, where the diagonal elements are the gammas for the correlated terms. The `CINITIAL` option is used to give initial values for this matrix. If no initial values are given, the initial values are taken from initial gamma values given in `VCOMPONENTS` when the model is declared. When correlations are declared between terms, you must set `FORMATION=whole`. In the random coefficient regression model above, no correlation structure is declared within terms since the subjects are independent. However, it is possible to declare correlation/covariance models within terms as usual.

### 3.4 Practical

Spreadsheet file `Wheat72.gsh` contains data from a trial of 35 varieties of wheat. The design has two replicates each laid out in a 5 by 7 plot array. Fit an  $AR1 \otimes AR1$  spatial model, and see whether there is any evidence of an additional replicate effect or of measurement error. Is there any advantage in increasing the order of either auto-regressive structure to  $AR2$ ?

### 3.5 The variogram

The *variogram* is a tool from Geostatistics, that can also be useful in spatial modelling (see the *Guide to the Genstat Command Language Part 2 Statistics*, Section 8.3). It assesses how the variance of the difference between observations at two points differs according to their spatial relationship. Mathematically

$$\text{variance}(x - y) = \text{variance}(x) + \text{variance}(y) - 2 \times \text{covariance}(x, y)$$

If we assume a stationary spatial process, the variances and covariances will depend only on the relative positions of the points, and not on their exact locations. Then

$$\text{variance}(x - y) = 2 \times \sigma^2 (1 - \rho)$$

where

$\sigma^2$  is the variance at point  $x$  and at point  $y$ , and

$\rho$  is the correlation between 2 points relatively spaced like  $x$  and  $y$ .

The *variogram* is the semi-variance, i.e.  $\frac{1}{2} \times \text{variance}(x - y)$  expressed as a function of their relative positions, and the *empirical* or *sample* variogram is an estimate of variogram made from the data. For example, we can estimate the semi-variance for points in adjacent rows by taking the  $\frac{1}{2} \times$  mean of the squared differences between the residuals at the pairs of data points that occur in adjacent rows.

Some typical variogram patterns are as follows:

- independent noise – sharp increase away from zero, otherwise flat;
- auto-regressive errors of order 1 with a high positive correlation – slow increase as lag increases, tailing off to constant;
- auto-regressive errors of order 1 with a low positive correlation – fast increase as lag increases, tailing off to constant;
- alternating pattern – usually reflects systematic row/column effects or may suggest negative correlation.

You may often see a mixture of patterns, for example AR1 + independent (measurement) error would generate a step change at zero followed by a smooth increase. However, interpretation can be rather subjective!

The two-dimensional variogram from a spatial model can be plotted automatically whenever you run a spatial analysis by checking the relevant box in the [Spatial Model Options](#) menu (Figure 3.5). Alternatively, you can click on the [Display variogram](#) button in the [Spatial Model Further Output](#) menu (Figure 3.9) to obtain the [Variogram Options](#) menu, which gives more control over the plot.

Or you can simply click on [OK](#) to use the default options.

**Figure 3.11**



The variogram from the final model in Section 3.2 is shown in Figure 3.12, and seems to be reasonably compatible with the anticipated pattern.

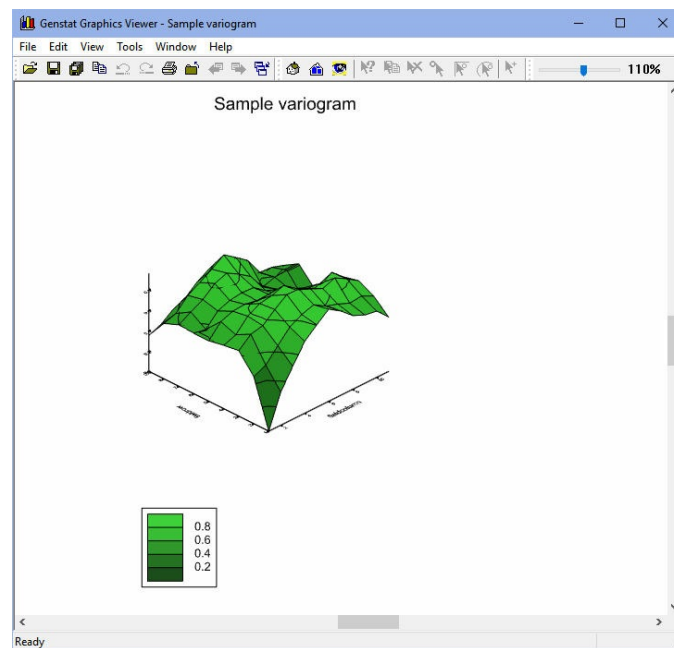


Figure 3.12

### 3.6 Practical

Plot and interpret the variograms for the model(s) fitted to the data in spreadsheet file *Wheat72.gsh* analysed in Practical 3.2.

### 3.7 Determining the random and correlation models automatically

The earlier sections of this chapter have explained how to use the spatial analysis menus to explore random models and correlation structures, and find the one that best explains the random variability. Strategies for doing this with incomplete-block and row-column designs have been programmed in Genstat procedures, and these are used by the automatic REML analysis menus, to enable you to determine appropriate random or correlation models automatically.

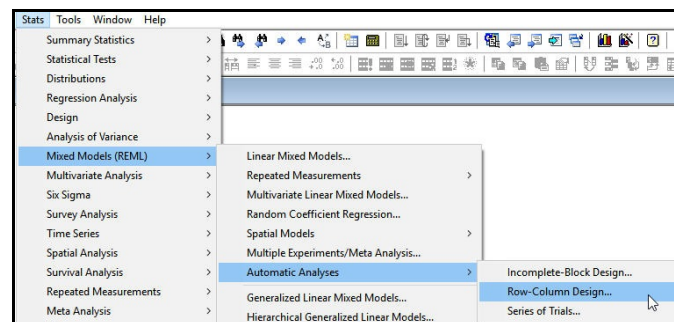


Figure 3.13

The ideas can be illustrated by the menu for *Automatic Spatial Analysis of Row-Column Designs* (Figure 3.14), which uses the procedure *VAROWCOLUMNDESIGN*. To open the menu, you click *Stats* on the menu bar, select *Mixed Models (REML)*, select *Automatic Analyses* and then click on *Row-Column Design* (see Figure 3.13).

To analyse the Slate-Hall farm data, we set the **Y-variate** to `yield`, the **Replicate factor** to `replicates`, the **Row factor** to `fieldrow`, the **Column factor** to `fieldcolumn`, the **Plot factor** to `plotnumber`, and the **Fixed model** to `variety`. We also include `replicates`, in the **Additional random terms**, to ensure that this is included in the random when (below) we search for the "best model".

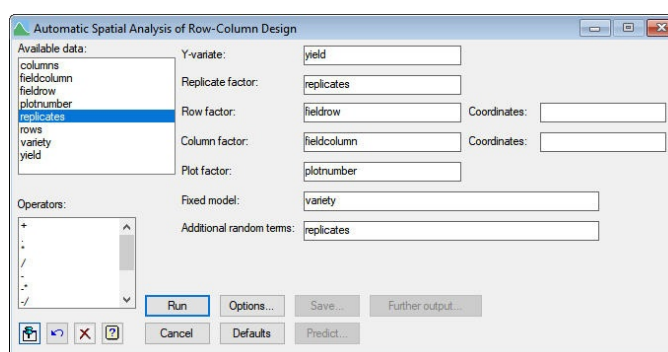


Figure 3.14

The **Y-variate**, **Row factor** and **Column factor** must be specified, but you do not need to supply a **Fixed model** or a **Replicate factor**.

The **Options** menu (Figure 3.15) controls the output, the models that are tried, and the strategy to use.

The **Display** box, at the top, controls the output that is produced about the models that have been tried.

The **Display output from best model** box allows you to display the most popular types of output from the best model. Others are available from the **Further Output** menu (Figure 3.16).

The **Model options** box controls several of the usual properties of the models, for example, whether to constrain variance components to be positive, or whether to include units with missing values. It also allows you to see whether linear trends are needed across the rows and the columns.

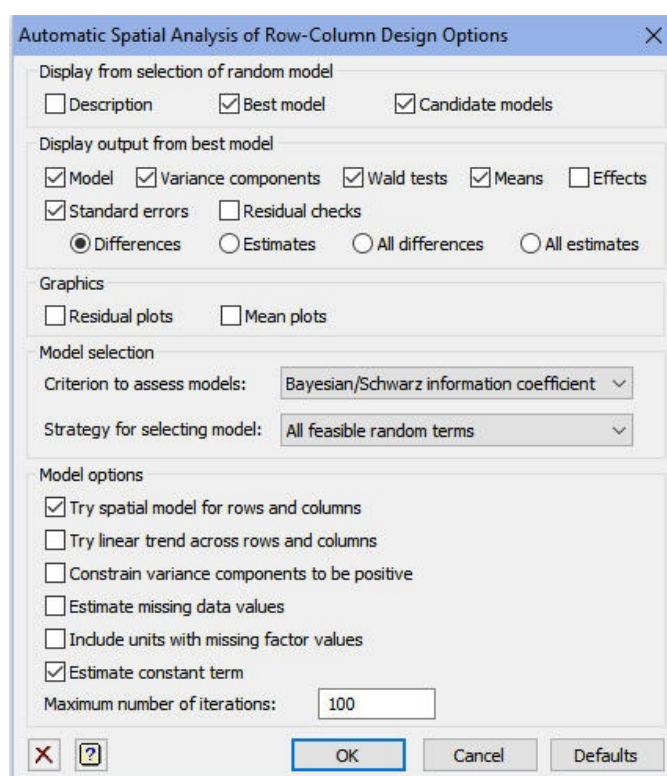


Figure 3.15

To do this, `VARCOLUMNDISEIGN` investigates whether two covariates are needed in the fixed model. These contain row and column coordinates respectively. The coordinates can be specified by entering variates into the **Coordinates** fields in the main menu. If these fields are left blank, the levels of the row or column factors are used instead. Finally, the **Model options** box allows you to choose whether or not to try to fit spatial models.

The strategy for selecting the random model is controlled by the **Model strategy** box. The **Criterion to assess models** list box indicates whether the models are to be assessed by their Akaike or Schwarz (Bayesian) information coefficients. The **Strategy for selecting model** list box specifies which random models are tried. The possibilities are as follows.

**All random terms**

fits the full random model, i.e. replicates, rows

	within replicates and columns within replicates if a replicate factor has been supplied, or rows and columns otherwise. This is appropriate if the row and column factors played a key role in the design and its randomization. For example, some factors may have been applied to complete rows or complete columns, as in a strip-block design.
All feasible random terms	tries to fit the full random model. If this is not possible, it tries models removing first one random term, then two and so on, until successful.
Only additional random terms	simply fits the additional random terms (if any). This is useful when you know the random model and want to investigate the effect of adding spatial covariance models.
Only feasible additional random terms	tries to fit the additional random terms. If this is not possible, it tries removing first one random term, then two and so on, until successful.
Best possible random model	tries all feasible random models. This may take a while, and so may be best left for the occasions when you are unsure what to do, or want to check the results from the fast strategy.
Fast strategy to find best model	follows an automatic strategy that aims to find the best random model without having to fit all of them. So, for example, it does not try models that include a column main effect as well as a spatial covariance model along rows.

The output below shows the results given by the menu settings in Figure 3.15. The first section of the output summarizes the models that have been tried by the fast strategy. There are several that we did not try earlier in this chapter, but the conclusion is the same. The best model (chosen as the one with smallest Schwarz Bayesian information coefficient) fits auto-regressive structures on the rows and columns, together with measurement error (with replicates as an additional random term). The analysis then reproduces the output shown in Section 3.2.

## Automatic REML analysis of row-by-column design

### Accumulated summary of REML random models

	Deviance	AIC	SIC	Random d.f.
No random terms	565.12	569.12	574.78	2
Rows	558.21	564.21	572.69	3
Columns	547.83	553.83	562.32	3
Rows & Columns	534.73	542.73	554.05	4
AR1(x)	540.12	546.12	554.60	3
Rows & AR1(x)	521.20	529.20	540.51	4
AR1(x)  & measurement_error	540.10	548.10	559.41	4

Rows & AR1(x)  & measurement_error	521.12	531.12	545.26	5
I(x)AR1	498.60	504.60	513.08	3
Columns & I(x)AR1	492.35	500.35	511.66	4
I(x)AR1 & measurement_error	498.59	506.59	517.90	4
Columns & I(x)AR1 & measurement_error	490.98	500.98	515.12	5
AR1(x)AR1	479.07	487.07	498.39	4
AR1(x)AR1 & measurement_error	471.93	481.93	496.08	5

Note: omits constant,  $-\log(\det(X'X))$ , that depends only on the fixed model.

Best model (by Schwarz Bayesian information coefficient):

AR1(x)AR1 & measurement\_error

## REML variance components analysis

Response variate: yield  
 Fixed model: Constant + variety  
 Random model: replicates + fieldrow.fieldcolumn + units  
 Number of units: 150

fieldrow.fieldcolumn used as residual term with covariance structure as below

Sparse algorithm with AI optimisation

## Covariance structures defined for random model

Covariance structures defined within terms:

Term	Factor	Model	Order	No. rows
fieldrow.fieldcolumn	fieldrow	Auto-regressive (+ scalar)	1	10
	fieldcolumn	Auto-regressive	1	15

## Estimated variance components

Random term	component	s.e.
replicates	-0.078	0.162
units	0.500	0.181

## Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
fieldrow.fieldcolumn			Sigma2	4.622	1.707
	fieldrow	AR(1)	phi_1	0.6922	0.1022
	fieldcolumn	AR(1)	phi_1	0.8442	0.0685

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	241.61	24	10.05	74.7	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
variety	241.61	24	10.05	74.7	<0.001

## Table of predicted means for Constant

14.47 Standard error: 0.894

## Table of predicted means for variety

variety	1	2	3	4	5	6	7	8
	12.45	15.14	14.05	14.06	14.71	15.21	13.74	14.53
variety	9	10	11	12	13	14	15	16
	12.63	11.95	13.29	14.40	16.24	13.00	14.69	12.86
variety	17	18	19	20	21	22	23	24
	14.94	15.28	16.50	16.44	15.18	16.10	13.18	15.55
variety	25							
	15.73							

Standard errors of differences

Average:	0.6072
Maximum:	0.6401
Minimum:	0.5719

Average variance of differences: 0.3689

---

The **Further Output** menu is similar to those earlier in this chapter. So you can print or plot additional information from the analysis of the best model, in the usual way.

Of course, no automatic method can guarantee to work as well as a detailed investigation by an expert. For example, to avoid the process becoming unmanageable, only first-order auto-regressive models are tried, and the trends that can be fitted are linear across the whole field.

So it may also be sensible to plot a variogram, to check how well the model has succeeded in representing the random variation (see Section 3.5). If you are unhappy with the results, the best model from the automatic menu should provide a good starting point for further investigation.

The menu for automatic analysis of incomplete-block designs (Figure 3.17) is very similar to the menu for row-column designs except that you specify a factor for blocks on the main menu, instead of factors for rows and columns.

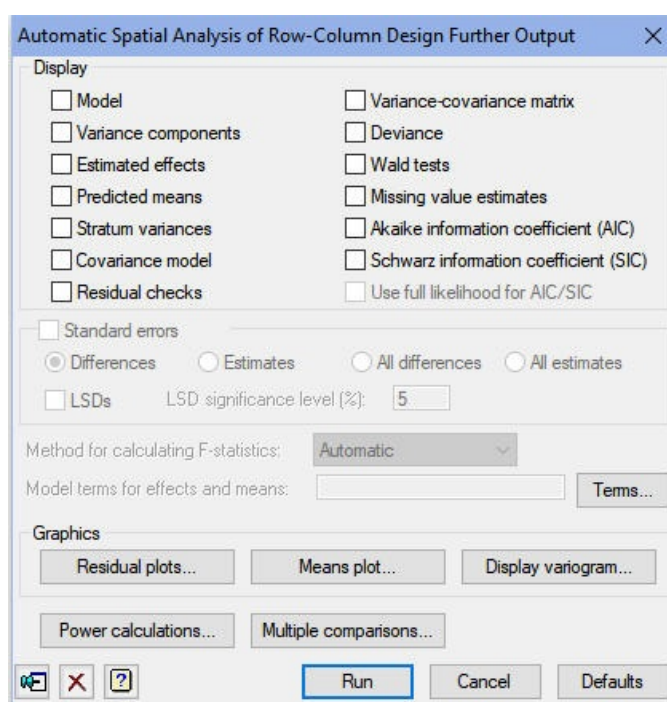


Figure 3.16

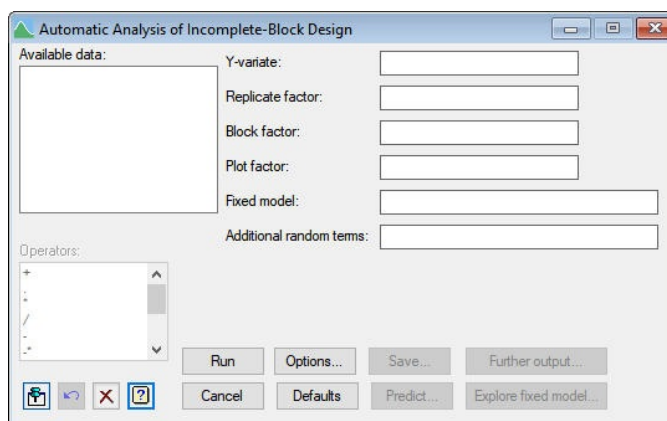


Figure 3.17



The factors for rows and columns are specified on the [Options](#) menu (Figure 3.18) if you want to fit correlation models.

**Figure 3.18**

### 3.8 Practical

Spreadsheet file `Gentrial2.gsh` contains data from a trial with 24 genotypes in a non-resolvable row-column design. Factors `Row` and `Column` specify the row and column coordinates of the plots. There is no replicate factor. (Therefore leave that box of the menu empty.) Some plots have missing values, and these must be estimated in order to retain the regular grid of plots. Analyse the trial using the best random model.

Why are no correlation models fitted across the columns? (Hint: how many columns are there.)



### 3.9 Automatic analysis of series of trials

The menu for **Automatic Analysis of Series of Trials** uses the **VASERIES** and **VAMETA** procedures to do a meta analysis of a series of trials with either row-column designs or incomplete-block designs (i.e. those that can be analysed by the other two automatic menus). You enter the block factor into the **Block factor** box, and the row and column factors are entered into

the **Row factor** and **Column factor** boxes. If all the trials have incomplete-block designs, the row and column factors need not be specified, and blocks need not be entered if they are all row-column designs. If there is a mixture, the row and column factors should either have only one level or missing values in each of the block designs, and the block factor should have only one level or missing values in each row-column design.

The spreadsheet file **SitesAtoD.gsh** (Figure 3.20) contains data from trials at four locations. There are 1310 new entries and two control genotypes. The controls are replicated in every trial. The new entries are unreplicated within the trials and do not occur in every trial. The entries must thus be assessed by the combined analysis of the results from all the trials. Missing values are again included to provide a regular grid.

The trials at locations A, B and D have row-column designs, while the trial at location C has an incomplete-block design. The block factor for trials A, B and D therefore only contain level one. Similarly the row and column factors at location C only contain

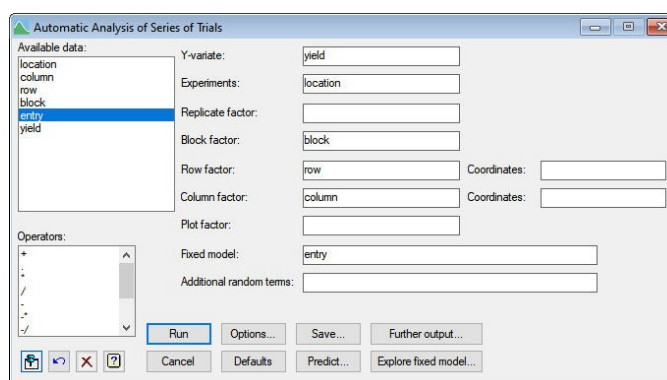
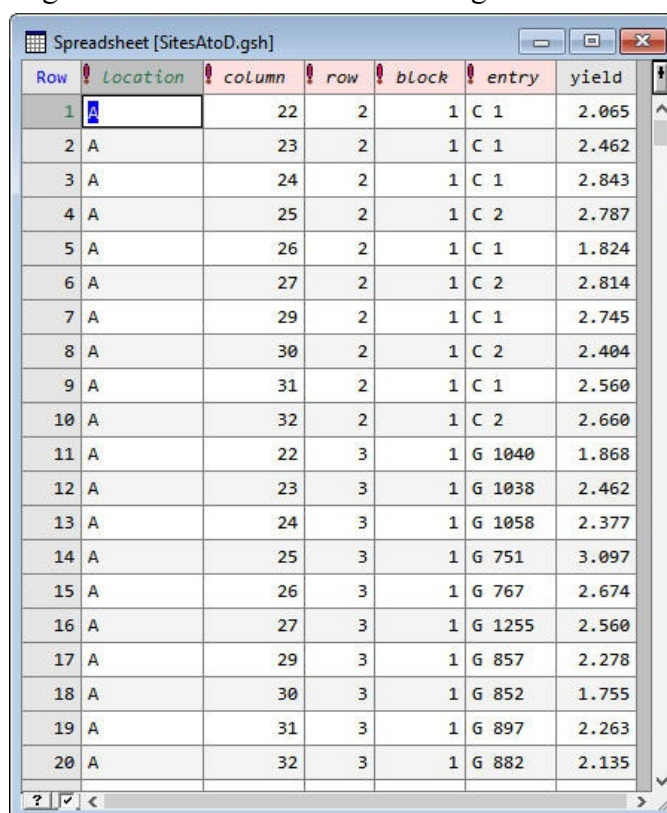


Figure 3.19

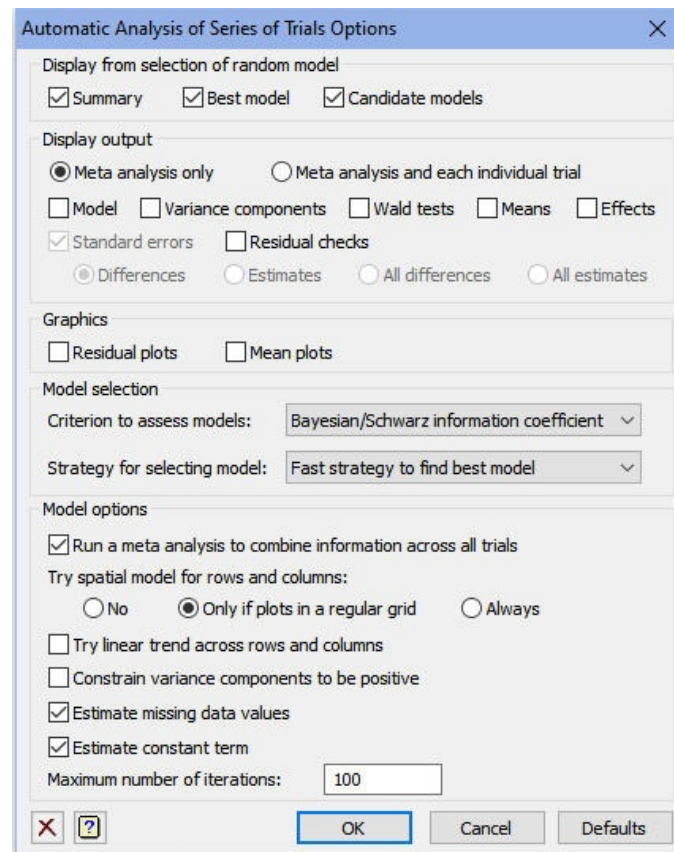


Row	location	column	row	block	entry	yield
1	A	22	2	1	C 1	2.065
2	A	23	2	1	C 1	2.462
3	A	24	2	1	C 1	2.843
4	A	25	2	1	C 2	2.787
5	A	26	2	1	C 1	1.824
6	A	27	2	1	C 2	2.814
7	A	29	2	1	C 1	2.745
8	A	30	2	1	C 2	2.404
9	A	31	2	1	C 1	2.560
10	A	32	2	1	C 2	2.660
11	A	22	3	1	G 1040	1.868
12	A	23	3	1	G 1038	2.462
13	A	24	3	1	G 1058	2.377
14	A	25	3	1	G 751	3.097
15	A	26	3	1	G 767	2.674
16	A	27	3	1	G 1255	2.560
17	A	29	3	1	G 857	2.278
18	A	30	3	1	G 852	1.755
19	A	31	3	1	G 897	2.263
20	A	32	3	1	G 882	2.135

Figure 3.20

In the options menu we have selected [Fast strategy to find best model](#) to reduce the amount of computation. We have checked [Summary](#) to print a summary of the models selected for the trials, [Candidate models](#) to see the models tried for each trial, and [Best model](#) to show the best one (chosen using the Schwarz Bayesian information criterion.) We have chosen not to display any other output. (We shall do this later using the further output menu.) We shall try spatial models only if the plots are in a regular grid (and estimate missing data values in order to achieve that). Finally, we have checked the box to run a meta analysis following the analyses of the individual trials.

The output is shown below.



**Figure 3.21**

## location A

## Automatic REML analysis of row-by-column design

### Accumulated summary of REML random models

	Deviance	AIC	SIC	Random d.f.
No random terms	43.16	45.16	47.33	1
Rows	43.12	47.12	51.46	2
Columns	37.12	41.12	45.47	2
Rows & Columns	37.12	43.12	49.64	3

Note: omits constant,  $-\log(\det(X'X))$ , that depends only on the fixed model.

Best model (by Schwarz Bayesian information coefficient): Columns

## location B

### Automatic REML analysis of row-by-column design

#### Accumulated summary of REML random models

	Deviance	AIC	SIC	Random d.f.
No random terms	161.18	163.18	165.81	1
Rows	161.00	165.00	170.25	2
Columns	152.81	156.81	162.06	2
Rows & Columns	139.98	145.98	153.85	3
AR1(x)	104.38	108.38	113.63	2
Rows & AR1(x)	103.52	109.52	117.39	3
AR1(x)  & measurement_error	103.98	109.98	117.85	3
I(x)AR1	106.70	110.70	115.95	2
Columns & I(x)AR1	104.23	110.23	118.10	3
I(x)AR1 & measurement_error	106.37	112.37	120.25	3
Columns & I(x)AR1 & measurement_error	103.98	111.98	122.48	4
AR1(x)AR1	75.32	81.32	89.19	3
AR1(x)AR1 & measurement_error	66.86	74.86	85.36	4

Note: omits constant,  $-\log(\det(X'X))$ , that depends only on the fixed model.

Note: table excludes models that could not be fitted successfully.

Best model (by Schwarz Bayesian information coefficient):

AR1(x)AR1 & measurement\_error

## location C

### Automatic REML analysis of incomplete-block design

#### Accumulated summary of REML random models

	Deviance	AIC	SIC	Random d.f.
No random terms	166.42	168.42	171.16	1
Blocks	158.08	162.08	167.57	2

Note: omits constant,  $-\log(\det(X'X))$ , that depends only on the fixed model.

Best model (by Schwarz Bayesian information coefficient): Blocks

location D

Automatic REML analysis of row-by-column design

Accumulated summary of REML random models

	Deviance	AIC	SIC	Random d.f.
No random terms	234.02	236.02	238.86	1
Rows	231.11	235.11	240.80	2
Columns	212.92	216.92	222.61	2
Rows & Columns	206.58	212.58	221.11	3
AR1(x)	167.19	171.19	176.88	2
Rows & AR1(x)	167.14	173.14	181.67	3
AR1(x)  & measurement_error	166.03	172.03	180.56	3
Rows & AR1(x)  & measurement_error	166.00	174.00	185.38	4
I(x)AR1	188.75	192.75	198.44	2
Columns & I(x)AR1	180.33	186.33	194.87	3
I(x)AR1 & measurement_error	188.59	194.59	203.12	3
Columns & I(x)AR1 & measurement_error	180.30	188.30	199.68	4
AR1(x)AR1	141.59	147.59	156.13	3
AR1(x)AR1 & measurement_error	138.92	146.92	158.29	4

Note: omits constant,  $-\log(\det(X'X))$ , that depends only on the fixed model.

Best model (by Schwarz Bayesian information coefficient): AR1(x)AR1

Summary: best model for each experiment

location	Best model
A	Columns
B	AR1(x)AR1 & measurement_error
C	Blocks
D	AR1(x)AR1

A different random model has been selected for every trial. The **VAMETA** procedure sets up the additional factors required to fit these in the meta analysis, in a similar way to the meta analysis in Chapter 2. These are labelled in the output by the name of the original factor, followed by @ and then the level or label of the trial: for example **block@C** for block effects in trial C.

You can use the [Automatic Analysis of Series of Trials Further Output](#) menu to display output from any of the individual trials or from the meta analysis, using the drop-down list at the top of the menu.

The output below shows the variance components, Wald tests and predicted means. Notice that, as a result of the complexity of the analysis, Genstat has been unable to estimate the denominator degrees of freedom to provide F tests for the fixed effect. The chi-square statistic is sufficiently large here for this not to be a problem!

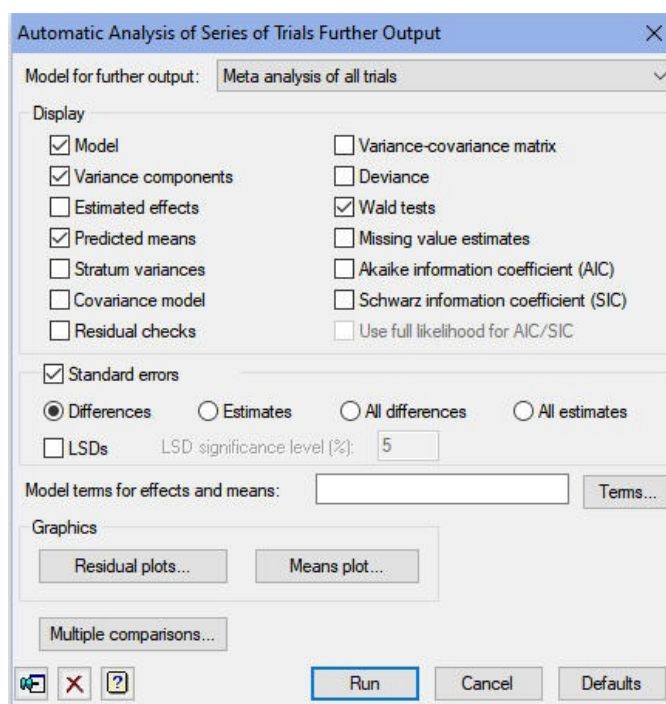


Figure 3.21

## Meta Analysis

### REML variance components analysis

Response variate: yield  
 Fixed model: Constant + entry  
 Random model: location + column@A + plots@B + block@C  
 Number of units: 2232

Separate residual terms for each level of experiment factor: location

Sparse algorithm with AI optimisation

Units with missing factor/covariate values included

- specific effect for term(s) omitted for units with missing values in column@A, plots@B, block@C

Units with missing data values included

### Residual models for multi-experiment analysis

Experiment factor: location

Experiment	Term	Factor	Model	Order	Nrows
A	Residual	Whole term	Identity	0	360
B	row.column	row	Auto-regressive	1	36
		column	Auto-regressive	1	18
C	Residual	Whole term	Identity	0	576
D	row.column	row	Auto-regressive	1	36
		column	Auto-regressive	1	18

## Estimated variance components

Random term	component	s.e.
location	1.8302	1.5112
column@A	-0.0009	0.0044
plots@B	0.0422	0.0118
block@C	0.0215	0.0105

## Residual model for each experiment

Experiment factor: location

Experiment	Term	Factor	Model(order)	Parameter	Estimate	s.e.
A	Residual		Identity	Variance	0.149	0.019
B	row.column			Variance	0.287	0.086
	row		AR(1)	phi_1	0.8990	0.0348
	column		AR(1)	phi_1	0.8308	0.0586
C	Residual		Identity	Variance	0.186	0.018
D	row.column			Variance	0.310	0.047
	row		AR(1)	phi_1	0.8332	0.0303
	column		AR(1)	phi_1	0.5013	0.0751

Warning 122, code VD 39, statement 1 on line 807

Command: VDISPLAY [PRINT=model,components,means,waldtests; PSE=Differences; FMET  
Error in AI algorithm when forming denominator DF for approximate F-tests.

## Wald tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	d.f.	Wald/d.f.	chi pr
entry	2553.59	1032	2.47	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	d.f.	Wald/d.f.	chi pr
entry	2553.59	1032	2.47	<0.001

*Message: chi-square distribution for Wald tests is an asymptotic approximation (i.e. for large samples) and underestimates the probabilities in other cases.*

*Message: negative variance components present. Fitting of fixed model terms is not sequential. Effects and means for any aliased fixed model terms may therefore be misleading. Wald tests, likelihood tests and fitted values are unaffected.*

### Table of predicted means for Constant

3.002

Standard errors of means are not available. Standard errors of differences can be obtained using PSE=diff or PSE=alldiff.

### Table of predicted means for entry

entry	G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8
	2.252	3.606	3.630	3.898	3.145	3.333	3.692	2.866
entry	G 9	G 10	G 11	G 12	G 13	G 14	G 15	G 16
	2.966	3.448	3.115	2.847	2.767	3.646	3.176	3.138
entry	G 17	G 18	G 19	G 20	G 21	G 22	G 23	G 24
	3.049	3.447	3.571	3.632	3.305	2.793	3.908	3.400
entry	G 25	G 26	G 27	G 28	G 29	G 30	G 31	G 32
	3.487	2.982	3.319	3.927	3.506	3.672	3.135	3.449
.....								
entry	G 1300	G 1301	G 1302	G 1303	G 1304	G 1305	G 1306	G 1307
	2.975	2.866	3.231	2.528	2.402	2.873	2.367	2.688
entry	G 1308	G 1309	G 1310	C 1	C 2			
	2.650	3.138	2.925	3.059	3.174			

Standard errors of differences

Average: 0.4412  
 Maximum: 0.6769  
 Minimum: 0.02945

Average variance of differences: 0.2014

## 3.10 Practical

Spreadsheet file [Gentrivals.gsh](#) contains data from a two trials with 24 genotypes in non-resolvable row-column designs. The trial at location 2 is the one analysed in Practical 3.8, and the trial at location 1 has a similar design. As before, factors [Row](#) and [Column](#) specify the row and column coordinates of the plots, and there is no replicate factor. Analyse the trials using the menu for [Automatic Analysis of Series of Trials](#).



---

## 4 Repeated measurements

The **REML** menus provide some very effective alternatives to the more traditional methods for the analysis of repeated measurements (see the *Guide to the Genstat Command Language Part 2 Statistics*, Section 8.1). There are parallels with some of the earlier methods. For example, Kenward (Applied Statistics, 1987) used ante-dependence structure to construct tests for changes in treatment effects between particular times and tests of treatment effects combined over all the times (*Guide Part 2*, Section 8.1.5), and this is one of the correlation models examined in Section 4.1. Likewise, random coefficient regression (Section 4.2) has similarities to the analysis of orthogonal polynomial coefficients calculated over time (*Guide Part 2*, Section 8.1.2). However, the **REML** framework provides a more flexible framework, in which you can compare one correlation model with another, or check the assumptions of the whole model.

In this chapter you will learn

- how to model covariances between time effects
- how to fit random-coefficient regression models ★

Note: the topics marked ★ are optional.

## 4.1 Correlation models over time

Figure 4.1 shows a Genstat spreadsheet, `ATP.gsh`, containing data from an experiment to study the effects of preserving liquids on the enzyme content of dog hearts. There were 23 hearts and two treatment factors, `A` and `B`, each at two levels. Measurements were made of ATP as a percentage of total enzyme in the heart, at one and two hourly intervals during a twelve-hour period following initial preservation. There is the choice of two menus for repeated measurements depending on whether the measurements are presented in separate variates (one for each time), or all in a single variate. The available analyses are identical, the menus merely provide different ways of specifying the data for the analysis.

Row	heart	time	A	B	ATP
1	1	0	1	1	85.51
2	1	1	1	1	74.56
3	1	2	1	1	84.25
4	1	3	1	1	84.05
5	1	4	1	1	78.41
6	1	5	1	1	79.93
7	1	6	1	1	71.6
8	1	8	1	1	69.9
9	1	10	1	1	63.52
10	1	12	1	1	57.14
11	2	0	1	1	76.54
12	2	1	1	1	72.77

Figure 4.1

Here the measurements are all in a single variate, `ATP`, and there are factors `heart` and `time` to indicate which heart provided the measurement in each unit of `ATP`, and the time when the measurement took place. Figure 4.2 shows the appropriate menu, which is obtained by clicking on **Stats** on the menu bar, selecting **Mixed Models (REML)** followed by **Repeated Measurements**, and then clicking on **Data in One Variate** (the alternative being **Data in Multiple Variates**).

Figure 4.2

In the figure we have filled in all the necessary boxes. Notice that we need to tell Genstat that the time points were the same for every subject (or heart). This would not be necessary if the measurements were in individual variates, one for each time.

The types of model that can be fitted differ according to whether the times of measurement were equally spaced or irregular. Auto-regressive and uniform correlation models can be fitted only to equally spaced measurements, whereas unstructured, antedependence or power models can be fitted in either situation.

In this example we shall fit ante-dependence models: a set of variates observed at successive times is said to have an ante-dependence structure of order  $r$  if each  $i$ th variate ( $i > r$ ), given the preceding  $r$ , is independent of all further preceding variates. In the analysis we start with order 1, and use the [Repeated Measurements Options](#) menu (Figure 4.3) to ask for only the deviance to be printed. The menu also sets the maximum number of iterations; in Figure 4.3 this is set to 30.

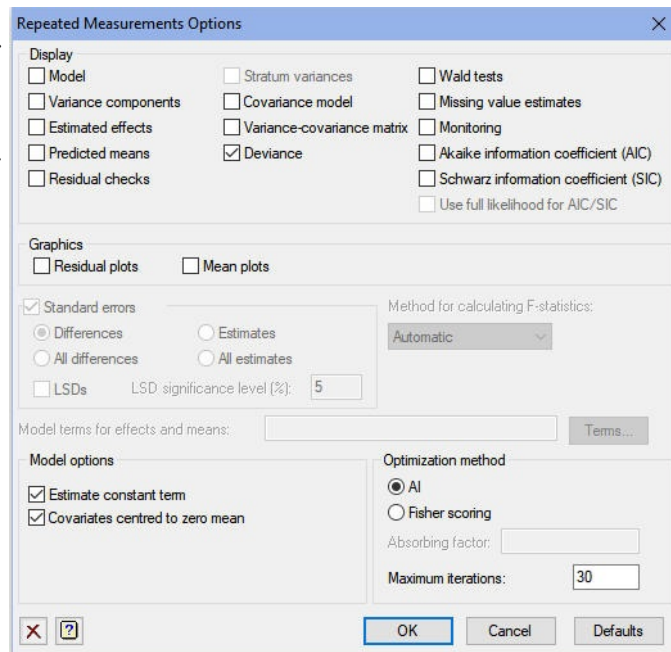


Figure 4.3

### Deviance: $-2 \times \text{Log-Likelihood}$

Deviance	d.f.
1021.84	171

Note: deviance omits constants which depend on fixed model fitted.

We can generalize the model by including additional uniform correlation within subjects (this is equivalent to including a random term for subjects, here the different hearts) or by increasing the order of ante-dependence to two. To investigate the first alternative we need to check the [Additional uniform correlation within subjects](#) box in Figure 4.2.

### Deviance: $-2 \times \text{Log-Likelihood}$

Deviance	d.f.
1011.50	170

Note: deviance omits constants which depend on fixed model fitted.

The change in deviance is 10.34. This is distributed as  $\chi^2$  on one degree of freedom. So there is definite evidence to support including uniform correlation within hearts. Now changing the ante-dependence structure to order two produces the deviance below.

## Deviance: -2\*Log-Likelihood

Deviance	d.f.
1005.49	162

Note: deviance omits constants which depend on fixed model fitted.

The  $\chi^2$  value, 6.00 on 8 degrees of freedom, is not significant. So we set the [Repeated Measurements Options](#) menu to print the model, variance components and Wald tests, and then refit with an ante-dependence of order one (retaining the additional uniform correlation within subjects).

## REML variance components analysis

Response variate: ATP  
 Fixed model: Constant + time + A + B + time.A + time.B + A.B + time.A.B  
 Random model: heart + heart.time  
 Number of units: 230

heart.time used as residual term with covariance structure as below

Sparse algorithm with AI optimisation

## Covariance structures defined for random model

Covariance structures defined within terms:

Term	Factor	Model	Order	No. rows
heart.time	heart	Identity	0	23
	time	Antedependence	1	10

## Estimated variance components

Random term	component	s.e.
heart	10.521	4.512

## Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
heart.time	heart time	Identity Antedependence(1)	Sigma2	1.000	fixed
			-	-	-
			dinv_1	0.09156	0.04175
			dinv_2	0.06305	0.03047
			dinv_3	0.03440	0.01241
			dinv_4	0.03129	0.01108
			dinv_5	0.01486	0.00543
			dinv_6	0.02841	0.01039

dinv_7	0.009922	0.003256
dinv_8	0.01197	0.00394
dinv_9	0.01024	0.00335
dinv_10	0.009345	0.003043
u_12	0.2208	0.4184
u_23	-0.03478	0.37496
u_34	-0.05258	0.26517
u_45	0.5696	0.3692
u_56	0.2383	0.1711
u_67	-0.5744	0.3831
u_78	-0.5977	0.2004
u_89	-0.4168	0.2088
u_910	-0.6001	0.2217

## Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
time	274.00	9	25.52	34.2	<0.001
A	0.42	1	0.42	18.6	0.525
B	0.29	1	0.29	18.6	0.598
time.A	38.60	9	3.60	34.2	0.003
time.B	23.01	9	2.14	34.2	0.052
A.B	5.76	1	5.76	18.6	0.027
time.A.B	3.04	9	0.28	34.2	0.975

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
time.A.B	3.04	9	0.28	34.2	0.975

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

---

The output shows evidence of time effects and of interactions involving time, A and B. So we finish by using the **Repeated Measurements Further Output** menu (Figure 4.4) to print A by B by time tables of predicted means.

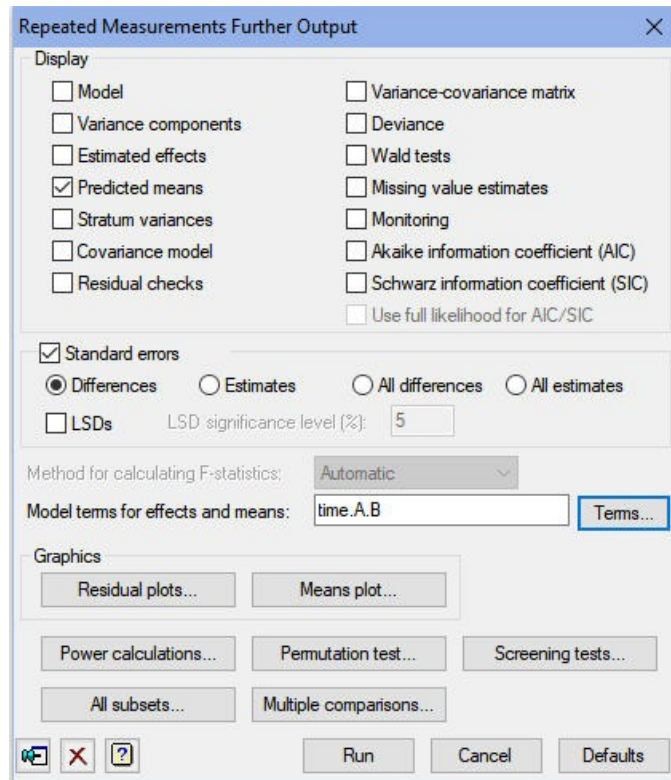


Figure 4.4

### Table of predicted means for time.A.B

	B	1	2
time	A		
0	1	77.47	84.14
	2	82.22	82.35
1	1	72.95	81.26
	2	84.38	81.94
2	1	79.31	82.74
	2	78.36	75.70
3	1	74.98	79.98
	2	75.16	75.48
4	1	76.10	75.17
	2	75.23	69.41
5	1	72.37	71.86
	2	73.46	61.59
6	1	64.38	61.33
	2	67.63	58.93
8	1	57.87	47.62
	2	68.92	56.73
10	1	48.40	43.68
	2	61.16	54.50
12	1	43.43	39.31
	2	56.83	51.52

## Standard errors of differences

Average: 5.088  
 Maximum: 7.670  
 Minimum: 2.314

Average variance of differences: 27.32

Standard error of differences for same level of factor:

	time	A	B
Average:	5.082	4.953	4.953
Maximum:	7.670	7.670	7.670
Minimum:	2.673	2.314	2.314

Average variance of differences:			
28.64	25.92	25.92	

To plot the means we click on the **Means Plot** button in the **Repeated Measurements Further Output** menu, to open the **Means Plot** menu shown in Figure 4.5. We specify **time** as the factor for the x-axis, and generate a trellis plot with **B** as the trellis factor, and **A** as the groups factor.

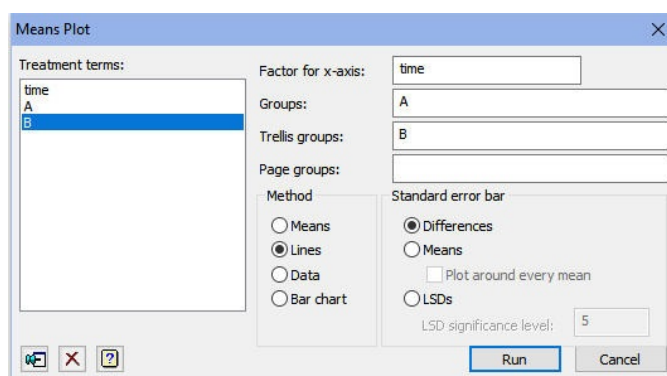


Figure 4.5

The plot (Figure 4.6) shows that the most interesting points on the **ATP** declines are just when the recording interval changed from hourly to every two hours!

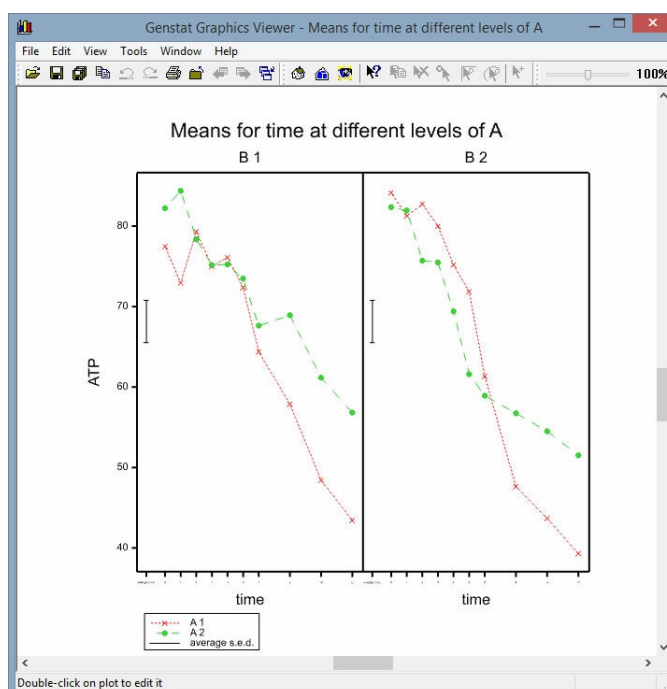


Figure 4.6



## 4.2 Practical

In a trial to study the effect of a dietary additive, seven rats were allocated at random to receive the standard diet, and seven to receive the enhanced diet. Their weight gains were measured after 1, 3, 5, 7 and 10 weeks. The data contained in Genstat spreadsheet `Ratmeasures.gsh`. Analyse the data using the **Repeated Measurements** menu. Hint: try an ante-dependence structure of order one with no random subject effects, or a power model, and remember that the data are in one variate.

## 4.3 Random coefficient regression

Random coefficient regression models the way in which the responses of the individual subjects change over time by fitting linear models with time as the x-variate. Often the model will be just a simple linear regression over time but in more complicated situations, polynomials may be used.

The analysis models all the coefficients of the polynomials for each subject simultaneously, assuming that each one comes from a random population whose means may depend on the treatments received by the subject. (The analysis is thus similar to the traditional analysis of polynomial contrasts over time, described in the *Guide to the Genstat Command Language Part 2 Statistics*, Section 8.1.2, but there only one polynomial coefficient is analysed at a time.)

Spreadsheet file `Boxrat.gsh` (Figure 4.7) contains the results of an experiment to study the effect of drugs on the growth rates of rats; see Box (1950, *Biometrics*, **6**, 362-389). There were three treatment groups: the first was a control, the second had thyroxin and the third had thiouracil added to their drinking water. The first column, `week`, is a factor recording the week when each weight was measured. The final column, `time`, contains the same information, but in a variate. The factor `rat` gives the number of the rat that was weighed in each measurement, and the factor `drug` indicates which treatment it received. The variate `weight` contains the weights.

Row	week	rat	drug	weight	time
1	1	1	control	57	0
2	0	2	control	60	0
3	0	3	control	52	0
4	0	4	control	49	0
5	0	5	control	56	0
6	0	6	control	46	0
7	0	7	control	51	0
8	0	8	control	63	0
9	0	9	control	49	0
10	0	10	control	57	0
11	0	11	thyroxin	59	0
12	0	12	thyroxin	54	0

Figure 4.7

We can plot the measurements using the **Repeated Measures Profile Plot - Data** menu (opened by selecting the **Profile Plot** sub-option of the **Repeated Measurements** option of the **Stats** menu on the menu bar). We click on the **Variate with time factor** button to indicate that the measurements are in a single variate with a factor to indicate the time at which each one was made. (The alternative is to have the data in separate variates, one for each time.) We then specify the data variate, and the time factor, and indicate that we would like separate lines to be plotted for each **drug**. See Figure 4.8.

The resulting plot, in Figure 4.9, shows that the profiles are mainly linear, with perhaps a hint of curvature. So we shall model them by quadratics.

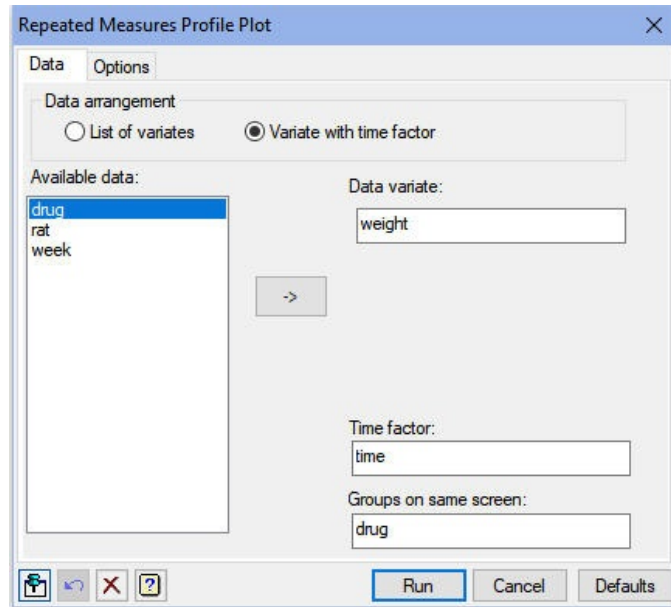


Figure 4.8

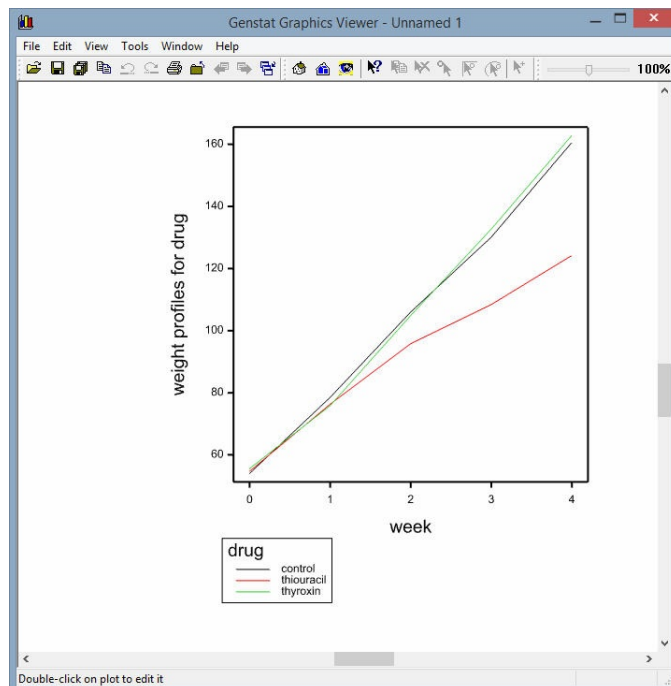


Figure 4.9

The **Random Coefficient Regression** menu (Figure 4.10) is opened by selecting the **Random Coefficient Regression** sub-option of the **Mixed Models (REML)** option of the **Stats** menu on the menu bar.

In the menu, we have defined **weight** as the data variate, **rat** as the subject factor, **time** as the variate of time points, and **timesq** as the identifier for the variate of squared time values (which will be calculated by the menu). The treatment structure is defined to be

```
drug * (time + timesq)
```

In the output, below, notice that **REML** has fitted covariances between the random intercepts and linear and quadratic regression coefficients from the random population of rats. This reflects the view that these are unlikely to vary independently from rat to rat – and this is confirmed by the estimated covariances that are obtained (see the parameters **v\_21**, **v\_31** and **v\_32**). The tests for fixed effects show no evidence of any differences overall between the drugs, but there may be differences in the linear and quadratic coefficients between the drugs.

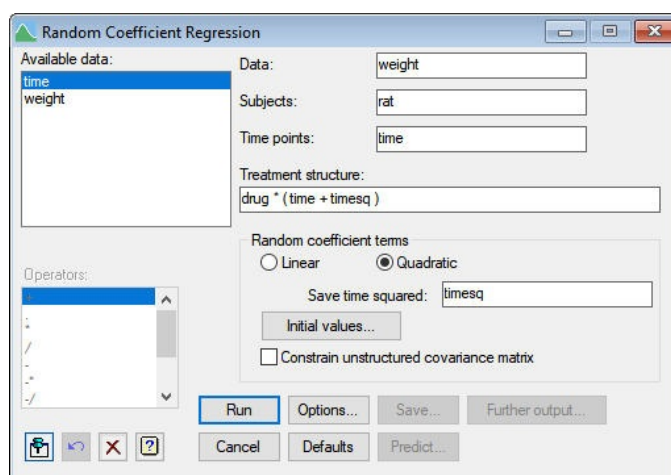


Figure 4.10

## REML variance components analysis

Response variate: weight  
 Fixed model: Constant + time + timesq + drug + time.drug + timesq.drug  
 Random model: rat + rat.time + rat.timesq  
 Number of units: 135

Residual term has been added to model

Sparse algorithm with AI optimisation  
 All covariates centred

## Covariance structures defined for random model

Correlated terms:

Set Correlation across terms  
 1 Unstructured

Set Terms  
 1 rat  
 1 rat.time  
 1 rat.timesq

Covariance model within term  
 Identity  
 Identity  
 Identity

### Estimated parameters for covariance models

Random term(s)	Factor	Model(order)	Parameter	Estimate	s.e.
rat + rat.time + rat.timesq	Across terms	Unstructured	v_11	8.653	3.072
			v_21	3.599	1.426
			v_22	1.504	0.953
			v_31	-0.2088	0.2661
			v_32	-0.2053	0.1997
			v_33	0.1113	0.0634
	Within terms	Identity	-	-	-
			-	-	-
			-	-	-
			-	-	-

Note: the covariance matrix for each term is calculated as  $G$  or  $R$  where  $\text{var}(y) = \text{Sigma2}(ZGZ' + R)$ , i.e. relative to the residual variance,  $\text{Sigma2}$ .

### Residual variance model

Term	Model(order)	Parameter	Estimate	s.e.
Residual	Identity	Sigma2	9.200	1.771

### Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
time	1094.47	1	1094.47	24.0	<0.001
timesq	0.11	1	0.11	24.0	0.745
drug	2.65	2	1.33	24.0	0.284
time.drug	21.66	2	10.83	24.0	<0.001
timesq.drug	19.85	2	9.93	24.0	<0.001

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
time.drug	0.91	2	0.46	24.0	0.640
timesq.drug	19.85	2	9.93	24.0	<0.001

*Message: denominator degrees of freedom for approximate F-tests are calculated using algebraic derivatives ignoring fixed/boundary/singular variance parameters.*

---

The menu conceals some fairly complicated programming, for example to obtain initial values for the covariance model. If you like more details about the process, you should read the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 5.4.5.

We can form predictions by clicking on the [Predict](#) button in the [Random Coefficient Regression](#) menu. We need to predict for both `time` and `timesq`, but notice that we can predict for `timesq` in “parallel” with `time` so that the resulting table has a single dimension for both of them; see Figure 4.11.

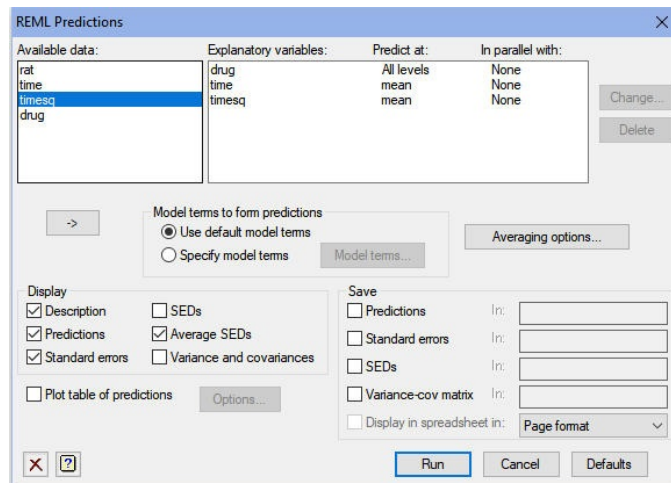


Figure 4.11

This parallelism can be set up when you use the [Change Prediction Values](#) menu to define the times at which you want to predict; see Figure 4.12. (The menu is opened by highlighting `time` in the [Explanatory variables](#) box, and clicking on the [Change](#) button.)

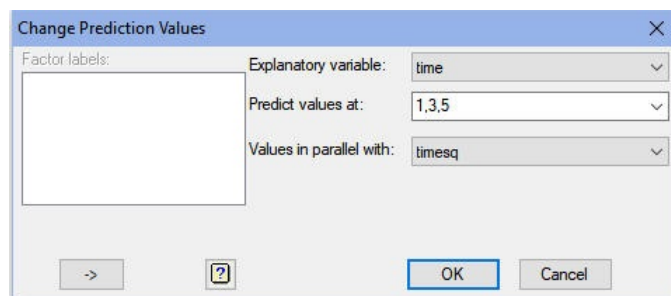


Figure 4.12

## Predictions from REML analysis

Model terms included for prediction: Constant + time + timesq + drug + time.drug + timesq.drug + rat.time + rat.timesq  
 Model terms excluded for prediction: rat

Status of model variables in prediction:

Variable	Type	Status
drug	factor	Classifies predictions
timesq	variate	Classifies predictions
time	variate	Classifies predictions
Constant	factor	Included in prediction
rat	factor	Averaged over - equal weights

Response variate: weight

Predictions

time_timesq	1.000 , 1.000	3.000 , 9.000	5.000 , 25.00
drug			
control	78.7	131.7	189.6
thiouracil	76.1	110.3	134.0
thyroxin	77.9	132.2	197.1

## Standard errors

time_timesq	1.000 , 1.000	3.000 , 9.000	5.000 , 25.00
drug			
control	2.576	3.533	6.137
thiouracil	2.576	3.533	6.137
thyroxin	3.010	4.346	7.760

Approximate average standard error of difference: 6.561 (calculated on variance scale)

---

#### 4.4 Practical

Spreadsheet file [PigGrowth.gsh](#) contains measurements made at weekly intervals in an experiment to study the effects of sex and diet on the growth of pigs. The experiment was in randomized blocks but, in your analysis you can treat this as a fixed term. Fit a random coefficient regression (with linear time only) to investigate the effects of sex and diet. Simplify the model to remove unnecessary terms. Investigate whether there are any quadratic time effects.

---

# Index

- \*units\* [71](#)
- Advantages of REML [1](#)
- Alpha design [7](#)
- Analysis of Variance menu [5](#), [10](#)
- Analysis-of-variance table [5](#), [42](#)
- ANOVA directive [7](#)
- Ante dependence model in REML [81](#)
- Ante-dependence [101](#), [102](#)
- AR1 model [66](#), [85](#)
- AR2 model [66](#)
- Auto-regressive [101](#)
- Auto-regressive model [66](#), [80](#)
- Auto-regressive moving-average model [80](#)
- Automatic analysis of incomplete-block design [91](#)
- Automatic analysis of row-column design [86](#)
- Automatic analysis of series of trials [93](#)
- Automatic search for the best random and correlation model [86](#)
- Banded covariance model for REML [81](#)
- BLUPs [31](#)
- Bounded linear covariance model
  - in REML [80](#)
- Circular covariance model
  - in REML [80](#)
- City block [77](#)
- Combined effects [44](#)
- Combined means [44](#)
- Conditional predictions [32](#)
- Conditional test [15](#)
- Correlation model [10](#), [66](#), [79](#)
  - automatic search [86](#)
  - over time [101](#)
- Correlations between model terms in REML [84](#)
- Cyclic design [7](#)
- Deviance [69](#), [70](#), [72](#)
  - in REML [102](#)
- Efficiency [42](#)
- Efficiency factor [42](#)
- Euclidean distance [77](#)
- Extracting results
  - from regression [49](#)
- F statistic [10](#)
- Factor analytic models for REML [81](#)
- Fertility trend [65](#)
- Field experiment [38](#), [63](#)
- Fixed effect [8](#)
- Fixed model [8](#)
- G matrix [66](#)
- Half-Normal plot [47](#)
- Information summary [42](#)
- Initial values for REML covariance models [81](#)
- Irregular layout
  - in REML [77](#)
- Keeping results
  - from regression [49](#)
- Lattice square [41](#)
- Least significant differences
  - for REML [23](#)
- Linear mixed model [2](#)
- Linear Mixed Model Further Output menu [11](#)
- Linear Mixed Model Options menu [8](#)
- Linear Mixed Model Save Options menu [49](#)
- Linear Mixed Models menu [8](#)
- Linear model [4](#), [65](#)
- Linear Regression Save Options menu [49](#)
- Linear trend [67](#)
- Linear variance model
  - in REML [81](#)
- Marginal test [14](#)
- Means
  - predicted [30](#)
- Means plot [26](#)
- Measurement error [71](#), [85](#)
- Meta analysis [51](#)
  - commands [59](#)
- Mixed Model [7](#)
- Moving average model for REML [80](#)
- Multi-level model [1](#), [2](#)
- Normal plot [47](#)
- Pedigree [82](#)
- Power distance covariance model for REML [80](#)
- Power model [101](#)
- Power-distance model [66](#)
- Prediction
  - from REML [36](#)
- Predictions [31](#), [111](#)
  - marginal [32](#)
- Profile plot [108](#)
- R matrix [66](#)
- Random coefficient regression [107](#)
- Random Coefficient Regression menu [109](#)
- Random effect [8](#)
- Random model [8](#)
  - comparing [69-72](#)
- Random term [67](#)
- Randomized-block design [3](#), [63](#)
- Regression
  - random coefficient [107](#)
- REML [2](#)
  - compared to ANOVA [45](#)
  - plotting means [27](#)
  - prediction from [36](#)
  - predictions [36](#)
- REML directive [20](#)
- REML Predictions menu [31](#), [111](#)



Repeated measurements [100](#), [101](#)  
Residual plots - from REML [48](#)  
Saving  
  regression results [49](#)  
  results to an external file [50](#)  
Screening tests [14](#), [15](#)  
Spherical covariance model  
  in REML [81](#)  
Spline [8](#)  
Split-plot design [3](#)  
Standard error  
  of difference of means [6](#)  
  of regression parameter [49](#)  
Standard errors  
  of differences of means [6](#)  
Storage  
  of results from regression [49](#)  
Stratum [5](#), [42](#)  
Stratum variance [44](#)  
Table of means [11](#)  
Traditional blocking [63](#)  
TREATMENTSTRUCTURE directive [7](#)  
Unequal variances in REML [82](#)  
Uniform correlation model [101](#)  
Unstructured model  
  in REML [101](#)  
Variance component [9](#), [19](#), [44](#)  
  constrain to be positive [73](#)  
  negative [73-75](#)  
Variogram [84-86](#)  
VCOMPONENTS directive [19](#)  
VDISPLAY directive [22](#)  
VGRAPH procedure [26](#), [27](#)  
VKEEP directive [49](#)  
VPLOT procedure [47](#), [48](#)  
VPREDICT directive [36](#)  
VRESIDUAL directive [59](#)  
VSPREADSHEET procedure [50](#)  
VSTRUCTURE directive [78](#)  
Wald statistic [10](#), [77](#), [103](#)  
Whole-plot [9](#)