NAG®

GENSTAT

Newsletter

Issue No. 29

Editors

P W Lane
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
HARPENDEN
Hertfordshire
United Kingdom AL5 2JQ

G W Morgan
NAG Ltd
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom OX2 8DR

Printed and Produced by NAG®
NAG is a registered trademark of:
The Numerical Algorithms Group Ltd
The Numerical Algorithms Group Inc
The Numerical Algorithms Group (Deutschland) GmbH
Genstat is a trademark of the Lawes Agricultural Trust
GLIM is a trademark of the Royal Statistical Society

ISSN 0269-0764

The views expressed in contributed articles are not necessarily those of the publishers.

Please note that the cover of this Newsletter has been adapted by kind permission of Oxford University Press, from the cover of the Genstat 5 Reference Manual.

Genstat Newsletter

Issue No. 29

NP2584

May 1993

Contents

		Page
1.	Editorial	3
2.	Genstat Talk	4
3.	A Note on Fitting a Growth-curve Model ... *M S Ridout*	6
4.	A Genstat Procedure for Fitting the Diggle-Zeger Model for Hormone Profiles ... *Roger P Littlejohn*	9
5.	Nonlinear Contrasts in ANOVA ... *R Butler and P Brain*	20
6.	Summaries of Unbalanced Factorial Data with Genstat ... *E D Schoen*	28
7.	Analysis of Overlapping Experiments with the REML Directives ... *G Horgan*	33
8.	A Method of Optimal Categorization of Discrete Variables ... *Paweł Krajewski*	41
9.	Handling Hierarchical Data ... *B M Church*	48
10.	Constructing a System within Genstat ... *J A Nelder*	50

Published by
The Rothamsted Experimental Station Statistics Department
and The Numerical Algorithms Group Ltd

Editorial

This Newsletter contains what is probably the last set of articles arising from the Seventh Genstat Conference at Papendal. In addition, among other contributions is a short summary of the K system by John Nelder: this is a GLIM-like interface to Genstat, based on procedures, which is available for purchase from NAG.

At the time of writing, two more Genstat conferences are about to take place. The Eighth Genstat Conference takes place at Canterbury, England, 19–23 July 1993. In addition, the first North American Genstat Workshop takes place in Kentville, Nova Scotia, 10–11 June 1993 after the Conference of the Statistical Society of Canada. This reflects the increasing interest in Genstat in Canada and the USA.

As usual, we invite people who are making presentations at these conferences to write articles for publication in this newsletter. The editorial team is about to change. Sue Welham at Rothamsted will be taking over from Peter Lane, who has been a co-editor since Issue 19 in 1987.

An electronic "list" has been set up for people to discuss Genstat. It is based at the Rutherford Laboratory in the UK, and is open to anyone who cares to join, from anywhere in the world. The intention is to provide a forum in which users of the system can exchange ideas, and to carry occasional news messages from Rothamsted and NAG. It is not intended as a bug-reporting mechanism: any problems with the system should still be reported direct to NAG, for example by email to **infodesk@nag.co.uk**.

To join the list, all you need to do is send the message
subscribe genstat name1 name2
to the address
listral@ib.rl.ac.uk (or **listral@uk.ac.rl.ib** for users in the UK) replacing **name1** and **name2** by your first and last names. Once you have joined, you can send messages to everyone on the list by mailing to
genstat@ib.rl.ac.uk

We hope to publish, in this newsletter, summaries of some of the subjects discussed on the list. Several have been prepared for this Issue as an experiment. Any views on this idea would be welcomed by the editors. We have not identified the discussants, because we cannot assume that they want their names publicized.

Genstat Talk

Genstat PC invocation of Genstat

Query: Why does the command line for running Genstat on the PC have such a non-standard form? Most packages happily accept a command like

```
  GENSTAT infile outfile
```

while Genstat seems to need

```
  GENSTAT in=infile,out=outfile
```

Also, many programs give interface hints in reply to a command like

```
  GENSTAT /?
```

Reply: The problem lies in the GENSTAT.BAT batch file. This performs a number of functions such as loading DBOS and setting up the environment for Genstat, and then calls Genstat with the desired parameters (such as IN=file). Unfortunately, DOS treats the = sign as a separator, so it is difficult to mimic satisfactorily both the forms with and without keywords. However, you can overcome this if you are prepared to do the initialization separate from the call to Genstat: for example, you could load DBOS and set the environment from the AUTOEXEC.BAT file. Then, you can dispense with GENSTAT.BAT and invoke the Genstat executable program directly: DOS will not then be involved with the interpretation of the command line. The "/?" idea will be considered for Release 3.

Genstat PC implementation requirements

Query: What are the minimum requirements for running Genstat on a PC, and are there separate versions for 386 and 486 machines?

Reply: The Installers' Note supplied by NAG with Genstat for PCs says: "Intel 80386 or 80486 based IBM or compatible PC with at least 2 Mb RAM. For 386 PCs a maths coprocessor (80387 or Weitek) is strongly recommended but not essential." The minimum installation needs 7 Mb disc space. My experience is that without a coprocessor on a 386, CPU time increases by a factor of 3 to 5. 486 machines have a coprocessor built in. There is no separate version for 486 machines, but they run Genstat faster because of the higher performance chip. As well as the 2 Mb RAM, it is advisable to use disk space as a 'swap area' if you intend to use more than the default data space (`S=1` in the Genstat command). This is explained in the Installers' Note.

Addendum: Only 486 machines with DX or DX2 chip have the coprocessor built in. An SX chip still requires one (usually on 20–33MHz machines).

Installing Genstat with a Smartdrive

Query: When installing Genstat I get a warning not to proceed because the installation will interfere with the smartdrive. Has anyone else had this problem, or know the way around it?

Reply: The installation instructions say that no virtual discs (e.g. VDISK) or disc caches (e.g. SMARTDRV) should be used because DBOS uses extended memory as program work area. If you wish to use a virtual disc, it must be instructed to leave enough extended memory (at least 2 Mb) for DBOS to load Genstat. They recommend that only the disc caching facilities built into DBOS should be used.

Genstat Versions 4 and 5

Query: Are there any utilities to help in the conversion of Genstat 4 programs to Genstat 5?

Reply: It was surprising that users did not contact NAG much for help in converting from Genstat 4 to Genstat 5. Perhaps they did not have much trouble, or perhaps they just started from scratch. There is help in the article 'Conversion from Genstat 4 to Genstat 5' in Newsletter 19. For further advice, it might be worth trying the discussion list!

Genstat and WordPerfect

Query: I am trying to get Genstat graphics from a PC using Device 2, into WordPerfect, but WP says it has an incorrect format. How do you do it?

Reply: The problem is that Device 2 in Genstat Release 2.2 on the PC produces simple PostScript, whereas WordPerfect 5.1 requires encapsulated PostScript. It is easy to get over this: just add the following two lines to the start of the PostScript file (which is a normal ASCII character file):

```
%!PS-Adobe-2.0
%%BoundingBox: 0 0 785 538
```

Addendum: A simpler solution is to use Device 4 or 5 to produce HPGL graphics. WordPerfect comes with a file called GRAPHCNV.EXE which can convert HPGL files to WordPerfect graphics format. A converted file can then be imported directly into WordPerfect, and can even be viewed on the screen unlike PostScript imports.

Survival data

Query: Has anyone any experience of fitting exponential and Weibull models to survival data and testing for goodness-of-fit in Genstat?

Reply: There are several examples in GLIM Newsletters and Aitkin *et al*'s book. Genstat can (as far as I know) fit any GLIM model.

Addendum: These models are easily fitted and tested in MLP. In Genstat, until the `DISTRIBUTION` directive becomes available with Release 3, you can use the `FITNONLINEAR` directive to fit distributions.

Generalized linear mixed models

Query: Has anyone handled generalized linear mixed models in Genstat?

Reply: There is a new procedure in Library 2[3] called `GLMM`, which was also described at the Genstat Conference at Rotorua. More presentations on this theme will follow at Canterbury.

Printing DOS Genstat graphs on HP lasers

Info: Genstat does not support a PCL device driver for the HP Laserjet 3, but does produce HPGL output. This will print on an HP laser if it is put into HPGL2 mode. However, on a network it is not feasible to do this manually, so I have written a small Turbo Pascal program to add the neccessary commands to the HPGL file to automate the process. (Files were enclosed, and are available from the discussion lists archive.)

REML

Query: Is there a limit (three perhaps) on the number of factors in a random effects term when using `REML`?

Reply: There is a bug in `REML` in Release 2. The `FACTORIAL` option is applied to both fixed and random terms, contrary to what is stated in the documentation. This will be fixed in Release 3; for the time being you can increase the setting of `FACTORIAL` beyond the default (3) to get round the problem. (This bug is described in error report E207 on the noticeboard: type

```
  NOTICE [errors]
```

to read the noticeboard.)

A Note on Fitting a Growth-curve Model

M S Ridout
Horticulture Research International
East Malling
WEST MALLING
Kent
United Kingdom   ME19 6BJ

1. Introduction

Plant growth is usually modelled using sigmoidal curves such as the Gompertz or logistic. Sometimes, however, interest lies only in the early stages of growth. This is true particularly of vegetable crops, which are usually harvested whilst their growth rate is still increasing. A model for the growth of vegetable crops that are receiving adequate supplies of water and nutrients is found to give a good fit to data from a wide range of crops. The model is defined by a simple differential equation that does not, however, have an explicit solution for dry weight as a function of time. The purpose of this note is to show that it is nonetheless easy to fit the model in Genstat, using the `FITNONLINEAR` directive.

2. The Model

Let w denote the dry weight of the plant and t denote time. The model is defined by the differential equation

$$\frac{dw}{dt} = \frac{K_2 w}{K_1 + w} \tag{1}$$

Thus, the rate of increase in plant dry matter is described by a rectangular hyperbola. In Greenwood *et al* (1977), the model is derived by assuming that the relationship between the net photosynthetic rate of the crop and its leaf area index is a rectangular hyperbola, and that the leaf area index is proportional to the dry weight of the crop.

It can be seen from equation (1) that when w is small, the relative growth rate is approximately K_2/K_1 and growth is close to exponential, whereas for large w the absolute growth rate is approximately K_2 and growth is close to linear.

Integrating equation (1) from some suitable timepoint, t_0, gives

$$w + K_1 \ln(w) = w_0 + K_1 \ln(w_0) + K_2(t - t_0) \tag{2}$$

where w_0 is the value of w at time t_0, but there is no explicit expression for w as a function of t. The simpler two-parameter model that arises when K_1 is constrained to be equal to 1.0 has been found to provide a good fit to data from many crops (Greenwood *et al* 1977).

3. Fitting the Model

We assume, as in Greenwood *et al* (1977), that it is appropriate to estimate parameters by applying the method of least squares to the (natural) logarithm of plant dry weight. This is appropriate if the variance of dry weight increases in proportion to the square of its mean, as is often observed with growth data.

Equation (2) can be re-written as

$$e^z + K_1 z = A \tag{3}$$

where $z = \ln(w)$ and A depends on t, t_0 and the unknown parameters w_0, K_1 and K_2.

Equation (3) is of the form $g(z) = A$ and a solution for z can be found by the iterative Newton–Raphson method in which the current estimate z_i is replaced by

$$z_{i+1} = z_i + \frac{A - g(z_i)}{g'(z_i)}.$$

In the present instance, this gives

$$z_{i+1} = \frac{e^{z_i}(z_i - 1) + A}{e^{z_i} + K_1}.$$

A simple initial estimate is

$$z_0 = \begin{cases} A/K_1 & \text{if } A \leq 0 \\ (A-1)/(1+K_1) & \text{if } 0 < A < 1. \\ \ln(A) & \text{if } A > 1 \end{cases}$$

Usually, each iteration of the Newton–Raphson algorithm is followed by a test for convergence to decide whether further iterations are necessary. For use with the **FITNONLINEAR** command, where the model is specified as a series of **CALCULATE** commands, it appears to be necessary to fix the number of iterations in advance. For the present model, numerical investigations show that, for any value of A, and for any value of K_1 in the range [0.1, 100], four iterations are sufficient to give an approximate solution to equation (3) with an absolute error less than 0.0001.

Since K_1 is often close to 1, it sensible to begin by fitting the constrained model with $K_1 = 1$. There are then simple procedures for obtaining initial estimates of K_2 and $\ln(w_0)$ which are outlined in the following section.

4. Example

The following data are measurements of dry matter (t/ha) of seedlings of wild cherry, Prunus avium, growing in a seed bed. The values are the mean dry weight of 15 seedlings, multiplied by an estimate of seedling density. Day number is counted from the beginning of the year. Previous applications of the model have been to vegetable and cereal crops and the experimenter was interested in whether the model might also describe the early growth of these tree seedlings.

```
UNITS [7]
READ [SERIAL=yes] DayNo,TotalDM
 135  168  188  207  230  251  274 :
0.09 0.76 2.37 5.12 6.22 8.37 11.5 :
```

The model is specified by three expressions. The first calculates A in terms of the current parameter values, the second calculates an initial estimate of $\ln(w)$ and the third is the equation for the iterative calculation of $\ln(w)$.

```
EXPRESSION Model[1...3]; VALUE= \
  !E(      A = EXP(LogW0) + K1*LogW0 + K2 * DayNo ), \
  !E( Fitted = (A/K1) * (A<=0) + LOG(ABS(A)) * (A>1) + \
               (A-1)/(1+K1) * (A>0 .AND. A<1) ), \
  !E( Fitted = (EXP(Fitted) * (Fitted-1)+A) / (EXP(Fitted)+K1) )
```

We begin by fitting the constrained model with $K_1 = 1$. An initial estimate of K_2 is then got by regressing $w + \ln(w)$ on $(t - t_0)$. Here t_0 is taken as the time at which the first measurement is made.

```
CALCULATE  YVar = TotalDM + (LogDM = LOG(TotalDM))
&         DayNo = DayNo - MIN(DayNo)
&       YVar, D = YVar, DayNo - MEAN(YVar, DayNo)
&        K2init = SUM(YVar * D) / SUM(D * D)
```

To get an initial estimate, u, of $\ln(w_0)$ the observed response at time t_0 is equated to its expected value.

```
CALCULATE u = (v = LogDM $[1]) * (v<=0) + LOG(ABS(v)) * \
(v>0) + 0.5 * (v-1) * (v>0 .AND. v<1)
FOR [NTIMES=4]
   CALCULATE u = (EXP(u) * (u-1) + v) / (EXP(u) + 1)
ENDFOR
```

Initial estimates of $K_2$and $\ln(w_0)$ are 0.117 and -2.49 respectively. The following four lines of code fit the constrained model with $K_1 = 1$.

```
MODEL LogDM; FITTEDVALUES=Fitted
SCALAR K1; VALUE=1
RCYCLE K2, LogW0; INITIAL=K2init, u
FITNONLINEAR [PRINT=m,s,e,f,c; CALCULATION=!P(Model[1,2,4(3)])]
```

Estimates of K_2 and $\ln(w_0)$ are respectively 0.108 (s.e. = 0.0068) and -2.52 (s.e. = 0.195). The residual sum of squares is 0.208 (5 d.f.).
The following two lines of code fit the unconstrained model

```
RCYCLE K1,K2,LogW0; INITIAL=1,K2init,u
FITNONLINEAR [PRINT=m,s,e,f,c; CALCULATION=!P(Model[1,2,4(3)])]
```

Estimates of K_1, K_2 and $\ln(w)$ are respectively 1.79 (s.e.=0.506), 0.149 (s.e. = 0.0259) and -2.44 (s.e. = 0.157). The residual sum of squares is 0.100 (4 d.f.). An approximate F-statistic for comparing the constrained and unconstrained models is therefore

$$4 * (0.208 - 0.100)/0.100 = 4.32 \text{ on 1 and 4 d.f.}$$

which gives a P-value of about 0.11. There is thus little cause to reject the simpler constrained model in this example.

5. Reference

Greenwood D J, Cleaver T J, Loquens S M H and Niedorf K B (1977) Relationship between plant weight and growing period for vegetable crops in the United Kingdom *Annals of Botany* **41** 987–997.

A Genstat Procedure for Fitting the Diggle-Zeger Model for Hormone Profiles

147

Roger P Littlejohn
AgResearch
Invermay Agricultural Centre
Private Bag
MOSGIEL
New Zealand

1. Introduction

I consult with scientists who measure hormones such as luteinizing hormone (LH) in sheep and deer. These hormones are released into the blood stream as a series of pulses, with concentrations rising to a sharp peak followed by a period of steady decay as the hormone is cleared from the system. Approaches used to analyse hormone profiles include pulse detection algorithms, spectral analysis and statistical modelling (Diggle and Zeger 1986). I generally use an adaptation (Littlejohn *et al* 1989) of the statistical model of Diggle and Zeger (1986, 1988) referred to here as DZ, in which instantaneous pulses are triggered by feedback with increasing probability as the hormone level decays. I have implemented this first-order Markov chain model as a Genstat procedure, details of which are given below, using data from Experiment 2 of Montgomery *et al* (1985) as an example.

2. The Diggle-Zeger Model

Denoting the hormone profile by $Y(t), t = 1...T$, the DZ model is specified by

$$Y(t) = \rho Y(t-1) + Z(t), \quad t = 2...T, 0 < \rho < 1,$$

where $Z(t)$ is independent of $Y(t-1)$ such that

i) $Z(t) \sim \Gamma(\mu, \nu)$ with probability $p(t)$
ii) $Z(t) \sim N(0, \sigma^2)$ with probability $1 - p(t)$, and
iii) $\log(p(t)/(1-p(t))) = \beta_0 + \beta_1 Y(t-1)$.

Thus, the $Y(t)$ consists of two components, the deterministic geometric decay from $Y(t-1)$ expressed by the autoregressive parameter ρ, and a random component which is a mixture of a gamma distributed random variable, modelling the release of a pulse, and a Normally distributed random variable with zero mean and variance σ^2, modelling sample and assay variability. The mixture probability $p(t)$, given by the logit regression on the previous level $Y(t-1)$, is the prior probability that an observation is a peak given the data up to and including $Y(t-1)$.

Littlejohn *et al* (1989) presented a variant on this, referred to as DZ$'$, with the conditions

i)$'$ $Z(t) \sim f_p(t) = LN(\mu, \tau)$ with probability $p(t)$
ii)$'$ $Z(t) \sim f_n(t) = N(0, Y(t-1)\sigma^2)$ with probability $1 - p(t)$, and
iii)$'$ $\log(p(t)/(1-p(t))) = \beta(\gamma - Y(t-1))$,

where $LN(.,.)$ is a lognormal random variable. The use of the lognormal peak amplitude is convenient in Genstat and τ is more stable to estimate than ν, although the gamma distribution is probably preferable in that, in common with the data, it has a shorter tail. Heteroscedasticity of noise variance was introduced because of greater variability at high hormone levels than later in the decay sequence, resulting in a tendency for large negative residuals to be associated with high fitted values. It was in fact necessary for convergence for some profiles in the dataset used here. The reparameterization of feedback reduces the correlation between parameters, and has the interpretation that γ estimates the median location at which feedback is triggered. It is then necessary to respecify the likelihood after iii) under the null hypothesis $\beta = 0$, if this test is required.

The posterior probability of a pulse at time t given the data up to and including $y(t)$ is given by

$$w(t) = p(t)f_p(t)/(p(t)f_p(t) + (1 - p(t))f_n(t)),$$

and referred to as the weight function. Substituting the parameter estimates into $w(t)$ gives a function that is usually close to either 1 or 0, interpreted as peaks or decay points, respectively.

One feature of hormone data that is not formally incorporated into the model is that some peaks rise over more than one sample, for example, at times 10–11, 25–26 and 31–32 of LH[3] in Appendix 1. This may be the case for 5–10% of points in a dataset. The model interprets these as two consecutive peaks, which is not the case. I will refer to such observations lying within an ascending sequence of values as "double rise points" and take the approach of Littlejohn *et al* (1989), of substituting each double rise point by the value of the previous nadir to calculate the probability that the next local maximum is a peak, while dropping the time of the double rise point from the likelihood.

The likelihood may then be written explicitly and maximized numerically. Details of the Genstat procedure code I use for this are given in Appendix 2 and enlarged upon in Section 3, and a discussion of the output is given in Section 4.

3. The Procedure DIGGLEZEGER

Appendix 2 contains three procedures: INITIALIZE, SWEEP and DIGGLEZEGER. DIGGLEZEGER sets up the likelihood function and carries out an analysis for each profile, firstly calling INITIALIZE to set the initial parameter estimates, then using SWEEP to detect and manipulate double rise points; initial values are reset using INITIALIZE. Then the negative log-likelihood is minimized using FITNONLINEAR. I routinely obtain further details of the analysis such as printouts and plots of the residuals and weight function, using a procedure named DIAGNOSTICS; the code is not included here, but output from it is included in Appendix 3.

Starting with DIGGLEZEGER, the option for monitoring FITNONLINEAR is set, and the log-likelihood is defined in expressions dz[1...13]. Points to notice here are that in dz[2,9]cutoff values of 70 and -8 are set to keep the calculations well-conditioned, while dz[4,6] preclude the possibility of a negative argument for the lognormal distribution and the variate double takes the value 1 for double rise points and 0 otherwise. The MODEL is then SAVEd in lsave to avoid being lost during INITIALIZE.

Each profile in turn is put into HORMONE and its first-order lag into HORMONE1 for analysis. Initial values for the parameters are calculated using moment estimators, assuming that any point for which the ratio $y(t)/y(t-1)$ is greater than QUOTLIM (=1.25 by default) is a peak. If no points satisfy this condition, analysis of that profile is aborted, with NJUMP set to zero. If only one point satisfies the condition, τ is initialized to a very small positive value. The likelihood and weight function are then evaluated at the initial parameter values to enable the assessment of double rise points. This is done in SWEEP, with REINIT set to 0 if there are no double rise points. Otherwise, HORMONE and HORMONE1 have double rise points reset as described in Section 2, and the location of all double rise points is put into double. If there are any double rise points, the initial parameter estimates are recalculated taking this into account. Natural parameter bounds and arbitrary steplengths are given via RCYCLE for calculating the maximum likelihood parameter estimates using FITNONLINEAR and the weight function.

At this stage, further diagnostics can be carried out using DIAGNOSTICS, which gives a set of summary statistics and (optional) data-residual-weight table and graphs. These depend on categorically assessing which points are peaks. An ambiguity arises for those points with weights between 0 and 1, so DIGGLEZEGER has the option WTLIM giving a cutoff weight for peak detection, with default setting 0.90. Thus a point with weight 0.85 would by default not be treated as a peak for obtaining summary statistics.

4. An Application

Appendix 3 contains the output of DIGGLEZEGER for LH[3]. Note that the initial parameter estimates for γ and β are very different from the final estimates, obtained after three rise points have been swept out. Other parameters have also been affected by this process. To obtain asymptotic standard errors the column denoted "sq. root of 2nd derivs" should be multipled by $\sqrt{2}$ (Lane 1991).

The summary statistics for this profile are based on a cutoff weight for peak detection of 0.90. For this profile all weights were either 0 or 1 to four decimal places. The difference between the "number of peaks" and the "sum of weights" gives some measure of how clear-cut the peaks are and how well the model fits the data. Peak amplitude statistics and fitted values are based on the lognormal parameter estimates, and the residuals are the difference between the fitted and observed values. The standardized residuals and increment relative to decay from the previous sample value (delta) are also printed out, to assist with the interpretation of peak detection. The plot of standardized noise residuals against fitted values shows a tendency for high fitted values to be associated with large negative residuals. The Normal scores plot for sorted noise residuals suggests that the model characterizes the variation satisfactorily for this profile.

5. Further Comments

I have subsequently generalized the model to incorporate double rise points in a less ad hoc manner than presented here. This uses a Fortran program that calls NAG optimization subroutines and is interfaced to Genstat using the OWN directive. Details will be given elsewhere.

6. Acknowledgements

I am grateful for a Trimble Fellowship which facilitated this research and enabled me to attend the 7th International Genstat Conference.

7. References

Diggle P G and Zeger S L (1986) Modelling endocrinological time series *ISI* **46** 1–12.

Diggle P G and Zeger S L (1988) A non-Gaussian model for time series with pulses *J. American Statistical Association* **84** 354–359.

Lane P W (1991) Minimization of a function *Genstat Newsletter* **27** 36–38.

Littlejohn R P, McWhirter J L, Henderson H V, Thompson J R, Montgomery G W and McMillan K L (1989) Analysing hormone profiles with pulses *The New Zealand Statistician* **24** 50–56.

Montgomery G W, Martin G B and Pelletier J (1985) Changes in pulsatile LH secretion after ovariectomy in Ile-de-France ewes in two seasons *J. Reproduction and Fertility* **73** 173–183.

Appendix 1. DZ.GEN

```
UNIT [NVALUES=37]
READ [PRINT=*] TIME,LH[1...10]
   1   6.6   6.1   3.1   2.6   8.5   4.8   2.8   3.4  10.1   3.5
   2   5.6   5.6   2.4   8.1   6.7   3.5   2.2   2.9   5.1   3.2
   3   4.3   4.6   1.9   5.5   5.4  13.0   2.1   6.5   4.1   3.1
   4   4.2   4.0   7.9   4.1   4.7   8.4   6.7   5.9   3.5   2.8
   5  12.0   3.4   5.7   3.5   4.2   6.3   4.8   4.2   2.8   1.8
   6   9.4  12.7   4.8   2.9   4.1   5.0   3.8   2.8   2.5   4.1
   7   8.0   6.8   3.6   2.5   8.3   4.3   2.7   2.8   2.0   4.1
   8   6.2   6.1   2.9   5.4   7.3   3.4   2.7   2.5   1.8   3.4
   9   5.0   5.4   2.1   7.2   6.0   2.7   2.0   7.5   5.5   3.2
  10   4.0   5.1   4.8   5.6   5.3  11.8   5.2   6.9   4.6   2.5
  11   3.1   5.1   6.4   3.9   4.5   8.4   7.3   5.3   3.8   2.0
  12   3.3  25.1   4.7   3.1   3.2   6.8   4.6   3.8   3.0   5.2
  13  14.1  11.1   3.9   2.4   7.3   5.3   3.6   2.9   2.5   3.8
  14  10.4   7.1   3.1   2.1   6.4   4.6   2.3   2.7   2.1   2.9
  15   8.1   7.1   2.6   2.1   5.0   3.7   1.9  15.7   1.8   2.7
  16   6.9   5.0   2.1  11.1   4.7   3.5  13.5   5.7   6.2   2.1
  17   5.4   9.1  14.1   6.7   4.3   4.2   7.4   4.2   5.0   1.8
  18   5.1  10.0   7.4   5.8   4.0  11.3   5.3   3.2   3.8   6.2
  19  28.5   8.1   6.7   4.3   3.3   8.7   4.6   2.7   3.2   4.8
  20  13.2   7.6   4.1   3.6   9.6   6.8   3.4   2.7   2.6   3.3
  21  10.2   6.9   3.9   3.9   9.4   5.3   2.7  19.5   2.3   2.8
  22   8.4   6.3   3.5   3.9   7.2   4.6  51.0   7.9   2.0   2.6
  23   6.7   5.4   2.6   3.7   6.1   3.8  10.5   5.3   7.3   2.1
  24   6.2  15.0   2.6  11.5   4.9   3.1   6.8   4.7   5.4   3.5
  25   4.7  10.2   7.3   9.5   4.5  18.5   5.2   4.3   4.1   6.5
  26  20.1   9.4   9.1   9.2   4.2   9.2   4.3   3.4   3.3   4.8
  27  13.2   7.8   7.7   6.3   8.7   6.6   3.5  29.5   2.7   4.0
  28  10.5   7.3   5.8   5.3   7.1   5.5  19.6   8.0   2.4   3.3
  29   8.3   4.7   4.4   4.5   5.5   4.6  11.6   6.0   2.4   2.6
  30   7.1   4.4   3.6   3.4   4.8   4.3   7.7   4.9   6.7   2.0
  31   6.2  11.1  10.3   2.9  15.1   3.2   5.7   5.1   5.5   8.7
  32   4.9  11.2  15.8  13.7  10.3  37.4   4.9   4.0   4.0   5.7
  33  22.1   7.8   8.3   5.7   7.3  10.4   4.0  19.8   3.2   4.4
  34  16.5   5.7   5.9   8.2   5.7   7.8   3.3   9.8   2.8   3.6
  35  10.4   6.2   5.2   6.0   5.2   6.9   2.3   7.7   2.5   2.8
  36   9.1   4.9   4.4   4.1   4.0   5.2  16.2   6.1   2.0   2.5
  37   7.3   3.8   3.3   3.4   3.7   4.7   7.7   4.2   8.7   7.9
   :

 DIGGLEZEGER [TPRINT=y; GPRINT=y; QUOTLIM=1.3; WTLIM=0.95] !P(LH[3])
 STOP
```

Appendix 2. DZ.PRC

```
JOB [INPRINT=s; DIAGNOSTIC=f,w] 'DIGGLE-ZEGER OPTIMIZATION'

PROCEDURE 'INITIALIZE'
  PARAMETER NAME='HORMONE','HORMONE1','QUOTLIM','DOUBLE','NJUMP', \
                 'GAMMA','BETA','MU','TAU','SIGMA','RHO'
  "Calculates naive initial parameter estimates"
  SCALAR beta0,beta1,ny
  CALCULATE ny=NVALUES(HORMONE)
  VARIATE [VALUES=#ny(1)] unity
  CALCULATE quotient=HORMONE/HORMONE1
  CALCULATE vquot=quotient.GE.#QUOTLIM
  RESTRICT quotient; (vquot.EQ.0) .AND. (DOUBLE.EQ.0)
  CALCULATE RHO=MEAN(quotient)
  CALCULATE SIGMA=SQRT(VAR(quotient))
[illegible]) .AND. (DOUBLE.EQ.0)
[illegible]-RHO*HORMONE1)
[illegible]
[illegible]
[illegible] in this profile'
[illegible]
[illegible])
[illegible]
[illegible]p)*(1-1/NJUMP))+0.0001*(NJUMP.EQ.1)
[illegible]
[illegible]E1; DOUBLE.EQ.0
[illegible]l] vquot; unity
[illegible]
[illegible]f
[illegible]1)
[illegible]
[illegible]
[illegible]E1
[illegible]


[illegible]
[illegible]ORMONE1','WEIGHT','DOUBLE','REINIT'
[illegible]ints and replaces them with nadir values"
[illegible]
[illegible]
[illegible]
[illegible]LACE(SHIFT(HORMONE; -1); 0).GT.HORMONE) \
[illegible]ONE.GT.HORMONE1)
[illegible]
[illegible]
[illegible]are detected, initialization is complete"
[illegible]
[illegible]
[illegible]
[illegible]
    EQUATE !P(mz,HORMONE); hormone0
    CALCULATE HORMONE=HORMONE-t3*(HORMONE-hormone0)
    EQUATE !P(mz,t3); t31
    CALCULATE t3=t3*t31
    EXIT [CONTROL=f] SUM(t3).EQ.0
  ENDFOR
  PRINT
  PRINT [IPRINT=*; SQUASH=y] \
      'There are double rise points in this profile which have been modified -'
  PRINT [IPRINT=*; SQUASH=y] \
      'The number of double rise points is ',dsum; DECIMALS=0
  CALCULATE HORMONE1=SHIFT(HORMONE; 1)
ENDPROCEDURE
```

```
PROCEDURE 'DIGGLEZEGER'
  OPTION NAME='MONITOR','TPRINT','GPRINT','QUOTLIM','WTLIM'; \
         MODE=t,t,t,v,v; DEFAULT='n','y','y',!(1.25),!(0.9)
  PARAMETER 'PROFILES'; MODE=p
  "Options control the printing of FITNONLINEAR, the table of fitted values
   and weights, and the diagnostic graphs;  and set cutoffs for the ratio for
   initial detection of peaks and the weight value for the final detection of
   peaks."
  TEXT [VALUES=s,e,c] fitpri,fitnprin
  IF MONITOR .EQS. 'y'
    DELETE [REDEFINE=y] fitnprin
    TEXT [NVALUES=4] fitnprin
    EQUATE !P(fitpri,!T('mon')); fitnprin
  ENDIF
  SCALAR ny,reinit,llike
  SCALAR gamma,beta,mu,tau,sigma,rho
  VARIATE unitct
  CALCULATE ny=NVALUES(unitct)
  VARIATE [VALUES=1...#ny] time
  "Specifies the likelihood for DZ'"
  "Feedback probability"
  EXPRESSION dz[1];  !E(phi=-beta*(gamma-hormone1))
  &          dz[2];  !E(feedback=1/(1+EXP(VMAX(!P(phi,!(#ny(-70))))))
  "Increment relative to decay from previous point"
  &          dz[3];  !E(delta=hormone-rho*hormone1)
  &          dz[4];  !E(deltap=(delta.LE.0) + delta*(delta.GT.0))
  "Peak density"
  &          dz[5];  !E(argp=(LOG(deltap)-mu)/tau)
  &          dz[6];  !E(fp=(delta.GT.0)*EXP(-(argp*argp/2))/(deltap*tau))
  &          dz[7];  !E(pfp=feedback*fp*(1-double))
  "Noise density"
  &          dz[8];  !E(stdres=delta/(sigma*hormone1))
  &          dz[9];  !E(argn=VMAX(!P(stdres,!(#ny(-8)))))
  &          dz[10]; !E(fn=EXP(-argn*argn/2)/(sigma*hormone1))
  &          dz[11]; !E(pfn=(1-feedback)*fn*(1-double)+double)
  "Log likelihood"
  &          dz[12]; !E(lpfppfn=LOG(pfp+pfn))
  &          dz[13]; !E(llike=-SUM(lpfppfn))
  MODEL [FUNCTION=llike; SAVE=lsave]
  SET [OUTPRINT=*]
  FOR hormone=PROFILES[]
    PAGE
    PRINT 'Optimization for profile ',!P(hormone)
    CALCULATE hormonei=hormone
    CALCULATE hormone1=SHIFT(hormone; 1)
    VARIATE [VALUES=#ny(0)] double
    "Initialize parameters assuming no double rise points"
    INITIALIZE hormone; hormone1; QUOTLIM; double; njump; gamma; beta; \
               mu; tau; sigma; rho
    "Abort profile if no sample value is > QUOTLIM x previous value"
    EXIT [CONTROL=f; REPEAT=y] njump.EQ.0
    PRINT [SQUASH=y] 'Initial parameter estimates:'
    PRINT [SQUASH=y] gamma,beta,mu,tau,sigma,rho; FIELD=10; DECIMALS=4
    SET [RSAVE=lsave]
    RCYCLE gamma,beta,mu,tau,sigma,rho; gamma,beta,mu,tau,sigma,rho; \
           gamma,beta,mu,tau,sigma,rho
    FITNONLINEAR [PRINT=*; CALCULATION=dz; NGRID=2]
    CALCULATE weight=pfp/(pfp+pfn)
    "Sweep out double points, if they exist"
    SWEEP hormone; hormone1; weight; double; reinit
    IF reinit.EQ.1
      "Reinitialize parameters with double rise points swept out."
      INITIALIZE hormone; hormone1; QUOTLIM; double; njump; gamma; beta; \
                 mu; tau; sigma; rho
      SET [RSAVE=lsave]
    ENDIF
    RCYCLE gamma,beta,mu,tau,sigma,rho; *,0,*,0,0,0; \
           *,*,*,*,*,1; .09,.04,.02,.015,.006,.0023
    FITNONLINEAR [PRINT=#fitnprin; CALCULATION=dz]
    CALCULATE weight=pfp/(pfp+pfn)
    "Carry out diagnostic procedures - code not given, but printout included:
    DIAGNOSTICS hormonei; hormone; hormone1; weight; stdres; delta; \
                mu; tau; rho; double; TPRINT; GPRINT; WTLIM "
    DELETE
  ENDFOR
ENDPROCEDURE
```

Appendix 3. DZ.LIS

```
 Optimization for profile       LH[3]

 Initial parameter estimates:
     gamma      beta        mu       tau     sigma       rho
    0.0908    0.3115    1.7332    0.5041    0.1365    0.7995

 There are double rise points in this profile which have been modified -
 The number of double rise points is           3


***** Results of optimization *****

*** Minimum function value: ***

      42.5364


*** Estimates of parameters ***
                                 sq. root of
                      estimate   2nd derivs
gamma                    2.471        0.420
beta                      2.34         1.62
mu                       2.091        0.247
tau                      0.391        0.175
sigma                   0.1061       0.0200
rho                     0.7837       0.0283


*** Scaled 2nd derivatives ***

estimate      ref     scaled 2nd derivatives

gamma           1     1.000
beta            2     0.363  1.000
mu              3     0.000  0.000  1.000
tau             4     0.000  0.000  0.000  1.000
sigma           5     0.000  0.000  0.000  0.000  1.000
rho             6     0.000  0.000 -0.036  0.013  0.003  1.000
                          1      2      3      4      5      6

 Mean hormone level for profile is                    5.35
 Number of peaks is                                      5
 Sum of weights is                                     5.0
 Expected peak amplitude is                           8.74
 Standard deviation of peak amplitude is              3.55
 Mean interval between peaks is                       7.00
 Standard deviation of interval between peaks is      1.41
 Coefficient of variation (%) is                      20.2


      time     LH[3]    double    fitted  residual     stdres     delta    weight
         1       3.1         *         *         *          *         *         *
         2       2.4         0       2.4     -0.03    -0.0899     -0.03    0.0000
         3       1.9         0       1.9      0.02     0.0748      0.02    0.0000
         4       7.9         0      10.2     -2.33    31.8144      6.41    1.0000
         5       5.7         0       6.2     -0.49    -0.5866     -0.49    0.0000
         6       4.8         0       4.5      0.33     0.5504      0.33    0.0000
         7       3.6         0       3.8     -0.16    -0.3181     -0.16    0.0000
         8       2.9         0       2.8      0.08     0.2058      0.08    0.0000
         9       2.1         0       2.3     -0.17    -0.5619     -0.17    0.0000
        10       2.1         1       1.6      0.45     2.0391      0.45    0.0000
        11       6.4         0      10.4     -3.99    21.3458      4.75    1.0000
        12       4.7         0       5.0     -0.32    -0.4654     -0.32    0.0000
        13       3.9         0       3.7      0.22     0.4342      0.22    0.0000
        14       3.1         0       3.1      0.04     0.1050      0.04    0.0000
        15       2.6         0       2.4      0.17     0.5184      0.17    0.0000
        16       2.1         0       2.0      0.06     0.2259      0.06    0.0000
        17      14.1         0      10.4      3.71    55.9182     12.45    1.0000
```

```
18       7.4       0      11.1    -3.65  -2.4412     -3.65   0.0000
19       6.7       0       5.8     0.90   1.1472      0.90   0.0000
20       4.1       0       5.3    -1.15  -1.6198     -1.15   0.0000
21       3.9       0       3.2     0.69   1.5792      0.69   0.0000
22       3.5       0       3.1     0.44   1.0721      0.44   0.0000
23       2.6       0       2.7    -0.14  -0.3854     -0.14   0.0000
24       2.6       0       2.0     0.56   2.0391      0.56   0.0000
25       2.6       1       2.0     0.56   2.0391      0.56   0.0000
26       9.1       0      10.8    -1.68  25.6112      7.06   1.0000
27       7.7       0       7.1     0.57   0.5886      0.57   0.0000
28       5.8       0       6.0    -0.23  -0.2875     -0.23   0.0000
29       4.4       0       4.5    -0.15  -0.2368     -0.15   0.0000
30       3.6       0       3.4     0.15   0.3248      0.15   0.0000
31       3.6       1       2.8     0.78   2.0391      0.78   0.0000
32      15.8       0      11.6     4.24  33.9924     12.98   1.0000
33       8.3       0      12.4    -4.08  -2.4366     -4.08   0.0000
34       5.9       0       6.5    -0.60  -0.6873     -0.60   0.0000
35       5.2       0       4.6     0.58   0.9205      0.58   0.0000
36       4.4       0       4.1     0.32   0.5886      0.32   0.0000
37       3.3       0       3.4    -0.15  -0.3181     -0.15   0.0000
```

```
     -+---------+---------+---------+---------+
     I     *                                  I
     I                                        I
 1.6 I        *                               I
     I                                        I
     I        *     *                         I
     I            *                           I
     I      *  ***      *                     I
     I     * * *                              I
 0.0 I     ** *                               I
     I       * 2 *   *                        I
     I      *      * *                        I
     I                *                       I
     I                                        I
     I                                        I
-1.6 I             *                          I
     I                                        I
     I                                        I
     I                             *  *       I
     I                                        I
     I                                        I
-3.2 I                                        I
     -+---------+---------+---------+----------
     0.        4.        8.       12.       16.
```

Std noise residuals vs Fitted values

```
     -+---------+---------+---------+---------+
     I    *      *     *        *     *       I
     I                                        I
 0.9 I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
 0.6 I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
 0.3 I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
     I                                        I
 0.0 I  ** ****** ***** ******** ***** *****  I
     -+---------+---------+---------+----------
     0.       10.       20.       30.       40.
```

Time series of weights

```
     -+---------+---------+---------+---------+
     I                                     *  I
     I                                        I
 1.6 I                                 *      I
     I                                        I
     I                             * *        I
     I                            *           I
     I                       *****            I
     I                     ***                I
 0.0 I                   *2                   I
     I               **2*                     I
     I            ***                         I
     I           *                            I
     I                                        I
     I                                        I
-1.6 I         *                              I
     I                                        I
     I                                        I
     I   *   *                                I
     I                                        I
     I                                        I
-3.2 I                                        I
     -+---------+---------+---------+----------
   -2.4      -1.2       0.0       1.2       2.4
```

Sorted noise residuals vs Normal scores

```
     -+---------+---------+---------+---------+
     I                                        I
     I                                .       I
 15. I                                ,       I
     I                 .              ,       I
     I                 ,              ,       I
     I                 ,             .,       I
     I                 ,             . .      I
     I                 ,.            . .      I
 10. I                 ,.            , .      I
     I                 ,.       .    , .      I
     I                . .       ,.   , .      I
     I    ,           . .      . ,   , ,      I
     I    ,.     .    .  ,     ,  .  ,  .     I
     I    ,.     ,.   .  ,     ,  . .   ,     I
  5. I    , .   . .   .   .    ,   ..    ,    I
     I   .      ,  .  .   ..  .    ,.     ,   I
     I . .   , .    . .     , .     ,      .  I
     I  ..    ,.     ,.      ,,               I
     I   ,                                    I
     I                                        I
  0. I                                        I
     -+---------+---------+---------+----------
     0.       10.       20.       30.       40.
```

Original time series

Nonlinear Contrasts in ANOVA

R Butler and P Brain
Department of Agricultural Sciences
University of Bristol
AFRC Institute of Arable Crops Research
Long Ashton Research Station
BRISTOL
United Kingdom BS18 9AF

1. Introduction

In many designed experiments, at least one of the factors involved has quantitative levels, and the aim of the experiment is often to see if there is a trend between the means for this factor, and whether this trend is influenced by other factors in the experiment. For example, an experiment was carried out to discover how the relationship between fresh weight and herbicide dose was altered by the volume of liquid in which the dose was applied. Six doses of a herbicide (10, 20, 40, 100, 160, 340 g/ha) were applied to a weed in two different volumes of liquid (Small, Large). Ten replicates were used, laid out in a randomized block design. The treatment means are shown in Fig 1. An analysis of variance of this data shows a strong interaction between dose and volume, but does not provide information about the changes in the dose–weed weight relationship. The graph shows clearly that weed weight decreases with increasing dose, so including this trend in the ANOVA would provide information as to how it was affected by volume.

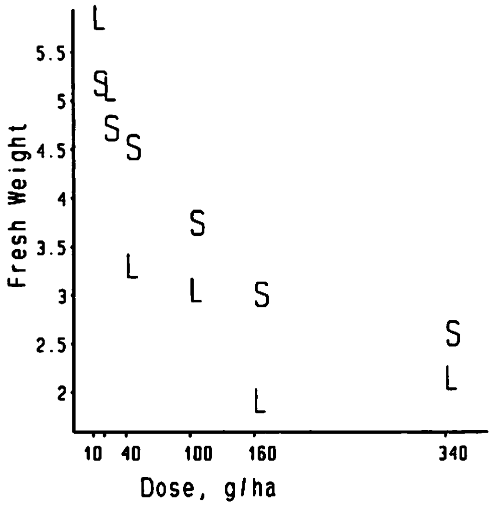

Fig 1 Example Treatment Means

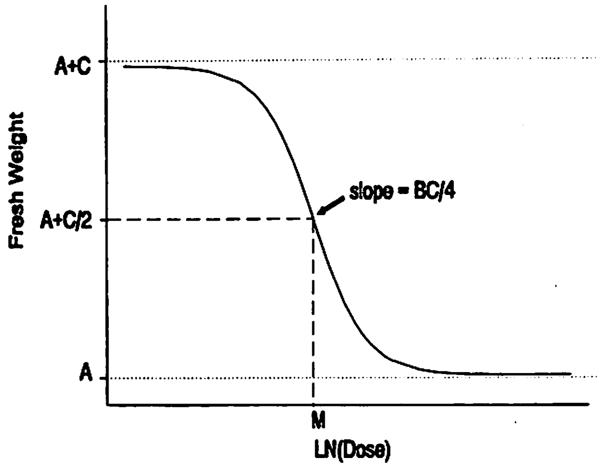

Fig 2 The Logistic Curve of log(Dose)

Genstat provides the **POL** function for use in the **TREATMENTSTRUCTURE** directive, which allows for the assessment of polynomial trends (contrasts) of many degrees; however, the fitting of these contrasts often has limited usefulness because they are usually not biologically meaningful. More usually, the trends of interest are nonlinear in character, but there is currently no direct way of assessing nonlinear contrasts within ANOVA. Quadratic trends could be fitted to the data in the example, but a more meaningful analysis would use the logistic curve of log(dose) to describe the relationship. This curve has parameters which have a direct biological interpretation (Fig 2). Fitting this curve instead of polynomials would give a more easily interpreted description of the effect of volume in this experiment.

A method has been developed to include the standard nonlinear curves of **FITCURVE** within the structure of ANOVA to allow such an assessment to be made.

2. Definitions

The measured variate (fresh weight in the example above) is denoted by y. The treatment structure consists of two factors: X, a quantitative factor with levels x, and another F, which has qualitative levels. In the example, X would be the Dose, with actual doses (or the log of the doses) in a variate x, and F would be Volume.

The POL command fits polynomial functions which can be written in the form:

$$y = a + \sum b_p x^p$$

where a and b_p are the parameters to be estimated. For the example, quadratic polynomial contrasts can be expressed in the following way:

$$\text{weed weight} = a + b_1 dose + b_2 dose^2$$

The parameters a and b_p may vary with differing levels of F (Volume).
The general form for a nonlinear function is

$$y = a + \sum b_p f_p(x; \theta_p)$$

where a, b_p, and θ_p are to be estimated, and which cannot be rewritten to make θ_p linear with respect to x. θ may be a list of nonlinear parameters. The logistic function described above has two nonlinear parameters B, and M, with the two linear parameters A and C.

$$\text{Fresh Weight} = A + \frac{C}{1 + e^{-B(\ln(dose) - M)}}$$

As for linear contrasts, the estimated parameters (a, b_p and θ_p) may vary with the levels of F.
There are often more than two treatment factors used in an experiment, but this article will be confined to cases with exactly two factors, only one of which is to be used with nonlinear contrasts.

3. Parallelism

With linear contrasts, three basic models can be fitted: single line, parallel lines and separate lines.

Model	**Form**	**Example**
1) Single line	$y = a + \sum b_p x^p$	FWt= $a + b_1 dose + b_2 dose^2$
2) Parallel lines	$y = a_i + \sum b_p x^p$	FWt= $a_i + b_1 dose + b_2 dose^2$
3) Separate slopes	$y = a_i + \sum b_{p_i} x^p$	FWt= $a_i + b_{1_i} dose + b_{2_i} dose^2$

(i refers to a level of F). Model 1 is part of the main effect of X; model 2 reflects a significant main effect of F. Model 3 is part of the interaction effect between F and X, and has separate a and b_p (linear) parameters for each level of F. For nonlinear models, there is a fourth model which is also part of the interaction, and for nonlinear contrasts model 3 can be referred to as 'Common Nonlinear', as θ is the same for all levels of F.

Model	**Form**	**Example**
1) Single Line	$y = a + \sum b_p f(x; \theta_p)$	FWt= $A + C/(1 + \exp(-B(\ln(dose) - M)))$
2) Parallel lines	$y = a_i + \sum b_p f(x; \theta_p)$	FWt= $A_i + C/(1 + \exp(-B(\ln(dose) - M)))$
3) Separate slopes (common nonlinear)	$y = a_i + \sum b_{p_i} f(x; \theta_p)$	FWt= $A_i + C_i/(1 + \exp(-B(\ln(dose) - M)))$
4) Separate curves	$y = a_i + \sum b_{p_i} f(x; \theta_{p_i})$	Fwt= $A_i + C_i/(1 + \exp(-Bi(\ln(dose) - M_i)))$

These different models are illustrated in Fig 3a and Fig 3b.

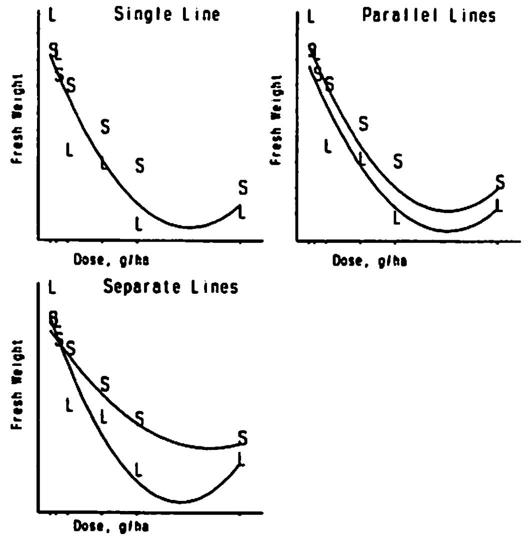

Fig 3a Fitted Quadratic Contrasts

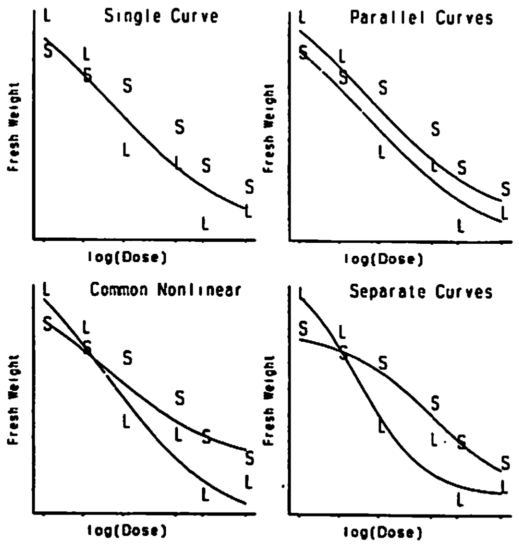

Fig 3b Fitted Nonlinear Contrasts

The sums of squares in the ANOVA table can be partitioned using these models, which allows an assessment of the importance of each as a component of the treatment effects. With polynomial contrasts, the sums of squares for the main effect of X (Doses) is partitioned into the SS for the single line, with the remaining SS as deviations. The Interaction SS is partitioned into SS for separate lines (adjusted for a single line), again with the remainder as deviations. Similarly, with nonlinear contrasts, the main effect is partitioned for a single curve and deviations, and the interaction for the common nonlinear curves (adjusted for a single curve), the extra SS for completely separate curves and deviations. The following outline ANOVA tables would be produced for the two cases:

Polynomial Contrasts

General Form	**Sum of Squares**	**Example**
F	SF	**Volume**
X	SX	**Dose**
Single Line		**Quadratic**
Deviations		**Deviations**
F.X	SFX	**Volume.Dose**
Separate Lines		**Volume.Quadratic**
F.Deviations		**Volume.Deviations**

Nonlinear Contrasts

General Form	**Sum of Squares**	**Example**
F	SF	**Volume**
X	SX	**Dose**
Single Curve	SSsingle	**Single Logistic**
Deviations	devX	**Deviations**
F.X	SFX	**Volume.Dose**
Common Nonlinear	SScommon	**Common Nonlinear Logistic**
Separate Nonlinear	SSseparate	**Separate Logistics**
F.Deviations	devFX	**Volume.Deviations**

4. Algebraic Derivation of Contrast Sums of Squares & DF

Analysis of Variance can be used to give the sums of squares for treatments (SF, SX, SFX). The sum of these is the total treatment SS (treatSS), which is the total SS for the (weighted) $F.X$ treatment means table. If the treatments are orthogonal to any blocks, then for $i = 1 \ldots t$ (levels of F), $j = 1 \ldots s$ (levels of X), and $k = 1 \ldots r$ (replicates),

$$\begin{aligned}
\mathrm{SF} &= \sum_i rs(\bar{y}_{i..} - \bar{y}_{...})^2 \\
\mathrm{SX} &= \sum_j rt(\bar{y}_{.i.} - \bar{y}_{...})^2 \\
\mathrm{SFX} &= \sum_{ij} r(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2.
\end{aligned}$$

Weighted regressions of the four nonlinear models on the treatment means gives four Residual SS:

	Residual SS
Single Curve	dev1
Parallel Curves	devP
Common Nonlinear	dev2
Separate Nonlinear	dev3

The form of these deviances can easily be written down using standard least-squares formulae. They each have the following form:

$$\text{deviance} = r\sum_{ij}(\bar{y}_{ij.} - \mathit{model})^2.$$

In each case, the constants a_i (or a) can be replaced by its least squares estimate.

Model:

1. $\hat{a} = \bar{y}_{...} - b\bar{f}_{.}$ $\quad \bar{f}_{.} = \underset{j}{\text{mean}}\,(f(x_j;\theta))$
2. $\hat{a}_i = \bar{y}_{i..} - b\bar{f}_{.}$
3. $\hat{a}_i = \bar{y}_{i..} - b_i\bar{f}_{.}$
4. $\hat{a}_i = \bar{y}_{i..} - b_i\bar{f}_{i.}$ $\quad \bar{f}_{i.} = \underset{j}{\text{mean}}\,(f(x_j;\theta_i))$

The deviances can be rewritten as follows:

a) dev1

$$\begin{aligned}
\mathit{dev1} &= r\sum_{ij}(\bar{y}_{ij.} - a - bf(x_j;\theta))^2 \\
&= r\sum_{ij}((\bar{y}_{ij.} - \bar{y}_{i..} - b(f(x_j;\theta) - \bar{f}_{.}))^2 + rs\sum_i(\bar{y}_{i..} - \bar{y}...)^2 \\
&= r\sum_{ij}((\bar{y}_{ij.} - \bar{y}_{i..} - b(f(x_j;\theta) - \bar{f}_{.}))^2 + SF \qquad (1)
\end{aligned}$$

b) devP

$$\begin{aligned}
\mathit{devP} &= r\sum_{ij}(\bar{y}_{ij.} - a_i - bf(x_j;\theta))^2 \\
&= r\sum_{ij}((\bar{y}_{ij.} - \bar{y}_{i..} - b(f(x_j;\theta) - \bar{f}_{.}))^2 \\
&= \mathit{dev1} - SF \qquad (2)
\end{aligned}$$

c) dev2

$$\begin{aligned}
\mathit{dev2} &= r\sum_{ij}(\bar{y}_{ij.} - a_i - b_if(x_j;\theta))^2 \\
&= r\sum_{ij}((\bar{y}_{ij.} - \bar{y}_{i..} - b_i(f(x_j;\theta) - \bar{f}_{.}))^2
\end{aligned}$$

d) dev3

$$dev3 = r\sum_{ij}(\bar{y}_{ij.} - a_i - b_i f(x_j;\theta_i))^2$$
$$= r\sum_{ij}((\bar{y}_{ij.} - \bar{y}_{i..} - b_i(f(x_j;\theta_i) - \bar{f}_{i.}))^2.$$

It can be seen from the above derivations that all the deviances involve deviations away from $y_{ij.} - y_{i..}$. In the case of a split-plot structure with F as the mainplot treatment,

$$y_{ijk} = \mu + \beta_i + \delta_j + \gamma_{ij} + \epsilon_{ik} + \eta_{ijk} \Rightarrow \bar{y}_{ij.} - \bar{y}_{i..} = \delta_j + \gamma_{ij} + (\bar{\eta}_{ij.} - \bar{\eta}_{i..}).$$

Thus, all the deviations involve only the sub-plot error, implying that in general, deviances from regressions on the means can be used to calculate SS for contrasts, providing that each treatment effect is estimated entirely within one stratum. This result holds for other designs.

The difference between deviances for two models gives the regression SS for the more complex model, adjusted for the simpler model. The regression SS for any model is the treatment SS minus the residual deviance. Equation (2) shows that it is not necessary to fit the parallel-curves model. The five required SS for the contrasts and deviations can therefore be derived using just SF, SX, SX, dev1, dev2, dev3 as follows:

a) SSsingle = TreatSS − dev1
= SF + SX + SFX − dev1

b) devX = SX − SSsingle
= dev1 − SF − SFX (from 1)

c) SScommon = devP − dev2
= dev1 − SF − dev2 (from 2)

d) SSseparate = dev2 − dev3

e) devFX = SFX − SSseparate − SScommon
= SFX − (dev2 − dev3) − (dev1 − SF − dev2)
= SF + SFX + dev3 − dev1

The diagram below (Fig 4) illustrates these calculations. In each case, the whole circle represents the total treatment SS.

The degrees of freedom can be derived in a similar manner.

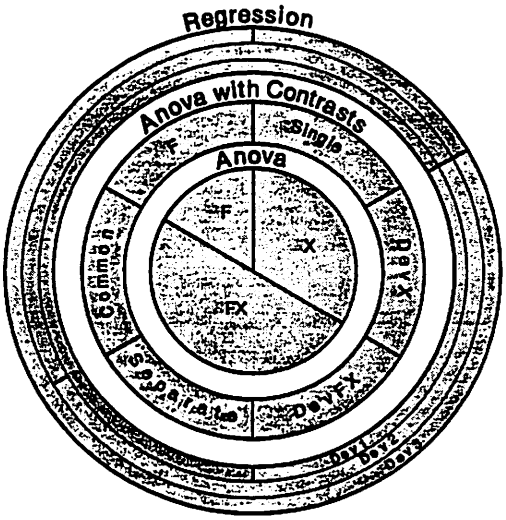

Fig 4 Partitioning Sums of Squares

5. Calculation Using Genstat

The above calculations for the derivation of the SS for contrasts can easily be carried out using the CALCULATE directive within Genstat, once the three models have been fitted using FITCURVE.

```
BLOCK block/plot/subplot         "Define block & treatment structures"
TREAT factor*varF
"Save means, reps & treatment SS from ANOVA"
ANOVA [PRINT=*] data
AKEEP varF.factor; MEAN=mean; REP=rep
& factor*varF; SS=s1,s2,s3

VARIATE [VALUES=#mean] Mean          "Put means & reps into variates"
&       [VALUES=#rep] Rep

CALC nf,nv=NLEVELS(factor,varF)"Create new factor & a variate for varF"
FACTOR [LEVELS=nf; VALUES=(1...nf)#nv] fac
VARIATE [VALUES=#nf(#variate)] variate

MODEL [WEIGHTS=Rep] Mean  "Fit the three models & save deviances & dfs"
TERMS variate*fac
FITCURVE variate                            "-single curve"
RKEEP DEVIANCE=dev1; DF=df1
ADD [PRINT=*] fac+variate.fac               "-Common Nonlinear"
RKEEP DEVIANCE=dev2; DF=df2
ADD [PRINT=*; NONLINEAR=separate]           "-Separate Nonlinear"
RKEEP DEVIANCE=dev3; DF=df3
```

CALCULATE can then be used to obtain the necessary SS and DF for the contrasts.
When the relevant numbers have been obtained, a new ANOVA table can be constructed. ADISPLAY is used save the ANOVA table in a text, and EDIT to add the extra lines.

6. A Procedure for Nonlinear Contrasts in ANOVA

The above method has been incorporated into the procedure NLCONTRASTS in the Genstat Procedure Library 2[3], which checks that the given design is a valid one for the fitting of nonlinear contrasts. The procedure has five options (PRINT, CURVE, FPROB, SE, WEIGHT), which are the same as the relevant options in ANOVA and FITCURVE. There are four input parameters: Y, XFACTOR (the factor to be used for contrasts), XLEVELS (values to be used for the levels of XFACTOR), and GROUPFACTOR, the factor whose interaction with XFACTOR is to be tested. Three more parameters (CONTRASTS, SECONTRASTS, DFCONTRASTS) save the contrast information in labelled pointers. The BLOCKSTRUCTURE and TREATMENTSTRUCTURE directives are used before the procedure is called as for a standard ANOVA, and the AKEEP directive can be used to save all the normal components of ANOVA, except those relating to contrasts. The TREATMENTSTRUCTURE can have terms with other factors, provided the main effects and interaction of XFACTOR and GROUPFACTOR are included.

As well as the analysis-of-variance table, and other standard output from ANOVA, the procedure also produces information about the contrasts fitted. Parameter estimates for each curve are given, along with standard errors for these based on the Residual MS for the stratum in which that contrast was fitted. Deviations between the fitted curve and the treatment means are available, each with standard errors. All nonstandard information can be saved as parameters for the procedure.

7. Example

The Genstat program below shows an analysis of the example data-set using NLCONTRASTS.

```
   1  UNITS [120]
   2  READ [CHANNEL=2] Fwt

    Identifier   Minimum      Mean   Maximum    Values   Missing
           Fwt     0.500     3.646     7.630       120         0

   3  VARIATE [VALUES=10, 20, 40, 100, 160, 340] dose; DECIMALS=0
   4  FACTOR [LABELS=!t(Small, Large); VALUES=60(1,2)] Vol
   5  FACTOR [LEVELS=dose; VALUES=10(#dose)2] Doses
   6  FACTOR [LEVELS=10; VALUES=(1...10)12] block
   7  FACTOR [LEVELS=12; VALUES=10(1...12)] pot
   8  CALCULATE ldose=LOG(dose)

  10  BLOCK block/pot
  11  TREAT Vol*Doses
  12  NLCONTRASTS [CURVE=logistic; PRINT=aov,con; FPROB=y] \
  13    Y=Fwt; XFACTOR=Doses; XLEVELS=ldose; GROUP=Vol

***** Analysis of variance *****

Variate: Fwt

Source of variation     d.f.       s.s.       m.s.    v.r.  F pr.

block stratum              9      4.580      0.509    0.43

block.pot stratum
Vol                        1      4.665      4.665    3.97  0.049
Doses                      5    163.164     32.633   27.78  <.001
  Curve                    3    159.314     53.105   45.21  <.001
  Deviations               2      3.849      1.925    1.64  0.199
Vol.Doses                  5     15.257      3.051    2.60  0.030
  Common NonLin            1      8.023      8.023    6.83  0.010
  Separate Curves          2      4.597      2.298    1.96  0.147
  Deviations               2      2.636      1.318    1.12  0.330
Residual                  99    116.287      1.175

Total                    119    303.952


***** Nonlinear Contrasts *****

*** Doses contrasts ***

 Parameter        Estimate         s.e.
 B                 -0.9272       0.7761
 M                   3.491       0.8472
 C                   4.961        3.390
 A                   1.686        1.083

 Deviations
        Doses           10         20        40         100        160        340
                    0.0153    -0.0726    0.1462     -0.2957     0.2718    -0.0651
           se       0.7822     0.9018    0.9295      0.9242     0.9614     0.7944


*** Vol.Doses contrasts ***

 Common Nonlinear
 Parameter        Estimate         s.e.
 B                  -1.013       0.4880
 M                   3.373       0.5204
 C  Vol Small        3.682
 A  Vol Small        2.484
 C  Vol Large        5.845
 A  Vol Large        1.292

 Separate Curves
 Parameter        Estimate         s.e.
 B  Vol Small       -1.269        1.291
 M  Vol Small        4.611       0.7709
 C  Vol Small        3.250        2.243
 A  Vol Small        1.875        1.664
 B  Vol Large       -1.703        1.009
 M  Vol Large        3.293       0.4123
 C  Vol Large        4.551        1.662
 A  Vol Large        1.961       0.3990
```

```
Deviations
           Doses        10        20        40       100       160       340
    Vol
  Small             0.0949   -0.1593    0.0446    0.1208   -0.1429    0.0419
             se     0.4303    0.8376    0.6070    0.7486    0.7063    0.1866
  Large            -0.0669    0.2023   -0.3107    0.5398   -0.3768    0.0123
             se     0.1576    0.4548    0.6228    0.7779    0.8816    0.5893
```

The ANOVA table shows that the significant interaction between Vol and Doses is due primarily to Doses affecting the slope parameter (C), whereas the nonlinear parameters (B,M) do not vary significantly with volume, since the 'Separate Curves' contrast was not significant. C represents the difference between the weed weight for zero dose, and A the lower asymptote. The parameter estimates show that this difference was significantly greater with a large volume, indicating a greater reduction in weed weight with herbicide sprayed as a large volume. This is illustrated in Fig 3b, 'Common Nonlinear'.

Summaries of Unbalanced Factorial Data with Genstat

E D Schoen
Centre for Applied Statistics
TPD-TNO
PO Box 6032
2600 JA DELFT
The Netherlands

1. Introduction

When factorial data are unbalanced, or when they are to be evaluated jointly with one or more explanatory variates, we should use the regression directives of Genstat to consider various statistical models. As a second step, we may wish to make adequate summaries of the results. Whenever factors come into view, a typical summary consists of tables in which the effects of one or more of the factors are averaged, possibly with extra dimensions classified by selected values of the variates in the model. To construct such summaries, the PREDICT directive should be used. It offers a wide range of possible methods for the averaging process. It is essential to choose the right options to guide the process, especially when the data are unbalanced. In particular, when not all combinations of factor levels are present in the data, or when there are aliased parameters, we should think carefully about the setting of the directive's options.

In the Genstat 5 Reference Manual, a rather compact illustration of each of the above options is given, necessarily without much discussion. The purpose of this presentation is to give some illustrations to facilitate the choice of the averaging process. I restrict my attention to linear models with Normal errors and constant variance.

2. Example 1: Additive Two-Way Model with Full Information on Parameters

The data given in Table 1 are classified by two factors called A and B. Each cell contains no more than one observation; for one reason or another, observations in the cells marked with an asterisk were not obtained. The column w_A gives the number of observations for each level of A; the row w_B shows the proportion of observations for the various levels of B.

The data fit perfectly in an additive structure. If the missing observations also fit in that structure, their values can be estimated. For example, the difference between the second and first level of B seems to be 2. This can be used to infer that the value of the upper left cell should be one. Table 1 can thus be completed. The full version is given as Table 2.

B		1	2	3	4	5	w_A
A							
1		*	3	*	*	*	1
2		2	4	6	6	13	5
3		10	12	14	*	21	4
	w_B	2	3	2	1	2	

Table 1. Data for example 1

B		1	2	3	4	5	w_A
A							
1		1	3	5	5	12	1
2		2	4	6	6	13	5
3		10	12	14	14	21	4
	w_B	2	3	2	1	2	

Table 2. Completed version of Table 1

The additive structure for the two-way table may be expressed with the well-known model

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

The model contains nine parameters: the general constant μ, one parameter for each level of A, α_i and one for each level of B, β_j.

For reading the data, fitting the additive model and printing the estimates of the parameters, the following lines of program suffice.

```
VARIATE [VALUES=3,2,4,6,6,13,10,12,14,21] Example
FACTOR [LEVELS=3; VALUES=1,5(2),4(3)] A
FACTOR [LEVELS=5; VALUES=2,1,2...5,1,2,3,5] B
MODEL Example
FIT [PRINT=estimates] A + B
```

This is the resulting output.

```
***** Regression Analysis *****

*** Estimates of regression coefficients ***

                               estimate        s.e.         t
Constant                        1.00000     0.00000         *
A 2                             1.00000     0.00000         *
A 3                             9.00000     0.00000         *
B 2                             2.00000     0.00000         *
B 3                             4.00000     0.00000         *
B 4                             4.00000     0.00000         *
B 5                            11.0000       0.0000         *
```

Genstat uses so-called corner-point constraints. For the data just shown this means that the first parameter of both of the main effects is set to zero. This implies that the fitted value of the upper left cell is given by the 'constant' in the output. (Of course, the table was artificially made completely additive. The standard errors of the estimates are therefore zero.)

When the model building with Genstat is done, we have to think of adequate summaries of the results. For a two-way layout, a typical summary would consist of tables classified by one or both of the classifying factors, possibly with extra dimensions classified by various values of the variates in the model. To construct such summaries, the PREDICT directive should be used. There are two parameters, namely CLASSIFY and LEVELS. With the latter parameter you specify for what levels of the factor you wish to obtain a summary, and for what values of the variates. By default, variates are evaluated at their mean value, and factors are evaluated at all their levels.

The CLASSIFY parameter lists the factors and variates that are to classify the summary table. The variates not listed with this parameter are evaluated at their mean value in the data-set. The factors not listed in the CLASSIFY parameter are averaged according to the setting of options. As we shall see below, it is quite important to choose the right option settings for averaging.

We may wish to base our summaries on the full reconstruction of the table. This is particularly interesting if it is perfectly possible to make observations at the combinations of factor levels now left empty and if you are quite sure of additivity there. This is how to get a full-table-based summary for factor B.

```
PREDICT [PRINT=predictions; COMBINATIONS=all; ADJUSTMENT=marginal] B
```

The statement produces the predictions given in the first column of Table 3.

```
COMBINATIONS                all                             present
ADJUSTMENT      marginal              equal         marginal            equal
 level
   1               5.100              4.333            5.556            6.000
   2               7.100              6.333            7.100            6.333
   3               9.100              8.333            9.556           10.000
   4               9.100              8.333            6.000            6.000
   5              16.100             15.333           16.556           17.000
```

Table 3. Summary tables for factor B made under various option settings of the PREDICT directive

The setting **all** of the **COMBINATIONS** option ensures that the summaries to be produced are based on Table 2 rather than on Table 1. It is the default setting of the option. The other one, **present**, produces summaries based on Table 1.

The **ADJUSTMENT** option controls the type of averaging. The setting **marginal** clearly does not yield simple averages over each of the columns. Instead, the program uses marginal weighting. The weights are just the total number of observations for each level of A. The value 5.1 in the table, therefore, is the sum of 1 times 1, 5 times 2 and 4 times 10, divided by the sum of the weights, 10 (see Table 2).

To see what happens it is illuminating to write down the formula used to calculate each entry in the one-way summary for factor B:

$$b_j = \hat{\mu} + \frac{1}{10}\hat{\alpha}_1 + \frac{5}{10}\hat{\alpha}_2 + \frac{4}{10}\hat{\alpha}_3 + \hat{\beta}_j.$$

Apparently, the difference between two elements of the summary table for B estimates the difference in the corresponding β parameters, and that is what we are frequently interested in.

Questions can be raised regarding the weighting. If the factor A represents a population trait, it is, in my opinion, perfectly sensible to use weights. If we know what the population proportion is, the obvious choice for the weights would be just these proportions. Genstat enables you to give weights explicitly with the **WEIGHTS** option.

Often enough you do not know what the population proportions are. In that case you may wish to use the obtained proportion as the weights. This is sensible if you could assume that the experimental units represent a random sample from the population of units. So for a complete data-set, the proportions 1 to 5 to 10 may have been obtained by random sampling. The use of these proportions for the weighting then gives a realistic picture of what happens if you give a randomly chosen experimental unit treatment B_i.

If the proportions are not related to a population, weighting is arbitrary. For example, we do not suspect that the proportion 1 to 5 to 4 in the data set of Table 1 represent a population proportion. Giving equal weights to the levels is then as good a choice as any. We can achieve this by setting the **ADJUSTMENT** option to **equal**. It gives a simple average over the rows or columns (see Table 3). The weights in the above formula all become 1/3.

You can easily see that comparison between cells in the one-way summary table still estimate the difference in β parameters. My own preference is to give equal weights unless there are population proportions to account for.

Both summaries on factor B discussed thus far are based on a reconstruction of the full data-set. If the empty cells in the table are due to 'impossible' factor-combinations, a reconstruction-based summary does not give realistic information on values which the response variate may take. With the setting **present** of the **COMBINATIONS** option, we ensure that only those combinations of factor levels are used that occur in the data. For the data at hand it results in quite spectacular changes (see Table 3). Particularly the value taken at the fourth level of B (only one value of the other factor used) is noteworthy.

For the setting **marginal** of the **ADJUSTMENT** option the summary for the various levels for B is calculated as follows

$$\begin{array}{lcccccccc}
b_1 = \hat{\mu} & & & + & \frac{5}{9}\hat{\alpha}_2 & + & \frac{4}{9}\hat{\alpha}_3 & + & \hat{\beta}_1 \\
b_2 = \hat{\mu} & + & \frac{1}{10}\hat{\alpha}_1 & + & \frac{5}{10}\hat{\alpha}_2 & + & \frac{4}{10}\hat{\alpha}_3 & + & \hat{\beta}_2 \\
b_3 = \hat{\mu} & & & + & \frac{5}{9}\hat{\alpha}_2 & + & \frac{4}{9}\hat{\alpha}_3 & + & \hat{\beta}_3 \\
b_4 = \hat{\mu} & & & + & \hat{\alpha}_2 & & & + & \hat{\beta}_4 \\
b_5 = \hat{\mu} & & & + & \frac{5}{9}\hat{\alpha}_2 & + & \frac{4}{9}\hat{\alpha}_3 & + & \hat{\beta}_5
\end{array}$$

It is evident that such summaries do not give you a clear idea about the difference between the βs in the model. So you do not get a clear picture about what happens if you change the level of factor B.

What you do win for the structural-empty-cell case is a realistic summary. More often than not, the weights given by **ADJUSTMENT = marginal** are arbitrary. So you may wish to use the **equal** setting.

3. Example 2: Additive Two-Way Model with Aliased Parameters

The data for the second example are given in Table 4. This data-set is even sparser than the one we studied before. Again, the table contains individual observations. In the upper left part of the table, additivity seems to hold. We cannot, however, reconstruct the whole table without assuming a value for the third parameter of factor A or of the fourth and fifth parameter of factor B. Indeed, these parameters are aliased.

B	1	2	3	4	5
A					
1	4	6	6	*	*
2	12	14	*	*	*
3	*	*	*	3	9

Table 4. Data for example 2

If we fit an additive model to the data using the regression directives of Genstat, the program decides that the aliased parameters are the last ones for the factor fitted last. They are set to zero. When it comes to predictions, the use of the `COMBINATIONS` option requires special attention. The default setting of this option, namely, `all`, will produce an error message. It is easy to see why. The default setting requires that the summary be based on a reconstruction of the whole table. Now any reconstruction of the lower left and upper right corners of the whole table must use the arbitrary value of the aliased parameter. The program is only prepared to do this if you give your explicit agreement. You do this by giving the `ALIASING` option of `PREDICT` the setting `ignore`. Genstat then uses the parameter values it already had in mind for the reconstruction of the whole table (the default setting of the option `ALIASING` is `fault`; this setting produces an error message when there are aliased parameters).

B	1	2	3	4	5
A					
1	4	6	6	3	9
2	12	14	14	11	17
3	4	6	6	3	9

Table 5. Reconstruction of Table 5 with estimate of α_3 set to 0

The table resulting from `PREDICT [ALIASING=ignore]` A,B is shown as Table 5. The only thing we have won by using the setting `ignore` is a reasonable reconstruction of the third cell of the second row. The upper right part of the table is also reconstructed, but the reconstruction is very arbitrary indeed!

Of course, it is not very tidy to analyse the data of Table 4 as an incomplete two-way factorial. I just wished to illustrate the handling of factorial data with aliased parameters.

One further potential application of the `ignore` setting may be mentioned. Let factor A have two levels, let factor B be quantitative with four levels, 16, 32, 48, and 64, say, and let a variate called 'linear' have the same values as factor B. Now the variate is equivalent to the linear part of the factor. If the factor is fitted first, you cannot add the variate to a regression model. So you should do it the other way round by stating for example:

```
FIT A*linear + B
```

If we want to predict the values of the response variable at the third level of B, say, we canot use

```
PREDICT A,B,linear; LEVELS=*,48,48
```

This would result in an error message because of the aliasing of B and linear. However, if we use the option `ALIASING=ignore`, the correct predictions will be formed.

4. Conclusions

In this paper, various ways of making summaries of factorial data with the PREDICT directive are discussed. We have seen that the directive offers quite a few possibilities for averaging. There seems to be no uniform best way of averaging; it all depends on the nature of the data at hand. A few personal recommendations are:

(1) If there are missing factor-combinations, we should judge whether they are structural or accidental. For accidental missing observations, we should base our summaries on a full table of predictions, including predictions for missing observations. For structural missing observations, we could base our summaries on just those combinations occurring in the data. We then probably need additional summaries which show what the effect of changes in factor levels are.

(2) If the data involve factors relating to population traits, we should weight the full table of predictions according to the population proportions. This is only sensible if we have at least some impression about these proportions. In all other cases we recommend equal weighting.

Analysis of Overlapping Experiments with the REML Directives

G Horgan
Scottish Agricultural Statistics Service
The University of Edinburgh
King's Buildings
EDINBURGH
United Kingdom   EH9 3JZ

1. Introduction

Data from designed experiments are analysed by analysis of variance. The Genstat directives for this technique (**BLOCKS**, **TREATMENTS**, **COVARIATE** and **ANOVA**) allow the data to be partitioned into several strata and a sophisticated analysis of variance to be constructed. Treatment means are currently estimated in the lowest stratum only — there are plans to change this. Only "balanced" designs can be analysed using these directives. Using pseudo-factors, it is possible to analyse designs that do not have full balance.

Residual Maximum Likelihood (Patterson and Thompson 1971, Robinson 1987) is a powerful technique when used for analysing data from unbalanced experiments. However, the technique must be used with care. The specification of the model is important, and there are aspects of this which do not arise in analysing experiments with **ANOVA**. This article will use data from three experiments, which when considered as a single experiment is unbalanced, to illustrate the use of REML, and to draw attention to some of the issues arising.

The essence of REML is that it maximizes the likelihood of all contrasts between the experimental observations which have expectation zero. It operates iteratively, and produces estimates of the variance components (sources of variability in the experiment — see Box *et al* 1978, p 571). The variance components are then used to produce weights for estimating treatment means, and so information from all strata is combined. REML has been available as a SASS (Scottish Agricultural Statistics Service) program since 1985, and the code has been incorporated in Genstat 5 since Release 2.1. The directive **VCOMPONENTS** specifies the model to be used, and the directive **REML** performs the analysis on specified variates.

2. Description of the Experiments

The three experiments were each concerned with examining the feed intake of pigs on different diets. Other variables, such as growth rate, were also recorded, but here we concentrate on the feed intake variable, recorded as g/day, averaged over the recording period for each pig. In the first experiment the pigs, in addition to their normal diets, were given three levels of a feed supplement (which we shall refer to as A). The three levels will be referred to as 0, 1, 2. The 0 level consisted of giving none of the supplement, and 1 and 2 were different amounts. Twelve pigs were used, arranged in four latin squares; the experiment was conducted over three periods. Two pigs followed each of the six treatment sequences.

	Pig					
	1&2	3&4	5&6	7&8	9&10	11&12
1	A0	A1	A2	A0	A2	A1
Period 2	A1	A2	A0	A1	A0	A2
3	A2	A0	A1	A2	A1	A0

The other two experiments were conducted in exactly the same way, using two other feed supplements, B and C.

3. Analysis

Considered as separate experiments, an analysis by ANOVA is straightforward, with BLOCKS pig*period and TREATMENT feed. A danger for the unwary, however, is that if the nine treatments are analysed as a single experiment, then inappropriate means are produced, and it is important that the warning Genstat produces is not ignored.

```
   1
   2  UNITS [108]
   3  FACTOR [LEVELS=6] TREAT
   4  FACTOR [LEVELS=3] period
   5  FACTOR [LEVELS=9; LABELS=!t(A0,A1,A2,B0,B1,B2,C0,C1,C2)] feed
   6  FACTOR [LEVELS=36] pig
   7  OPEN 'pigint.dat'; CHANNEL=2;
   8  READ [CHANNEL=2] pig,period,feed,feedint

    Identifier   Minimum      Mean  Maximum    Values   Missing
       feedint        22      1093     2714       108         0
   9  TABULATE [PRINT=means,nobs; CLASS=feed] feedint

                 Nobservd        Mean
         feed
           A0          12        1371
           A1          12        1616
           A2          12        1454
           B0          12        1169
           B1          12        1097
           B2          12         524
           C0          12        1149
           C1          12        1013
           C2          12         445

  10  BLOCK pig*period
  11  TREAT feed
  12  ANOVA feedint

******** Warning (Code AN 17). Statement 1 on Line 12
Command: ANOVA feedint

Partial confounding
feed is partially confounded with pig


12.........................................................

***** Analysis of variance *****

Variate: feedint

Source of variation           d.f.        s.s.        m.s.    v.r.

pig stratum
feed                             8    8166567.    1020821.   12.80
Residual                        27    2153215.      79749.    1.44

period stratum                   2    7300968.    3650484.   65.88

pig.period stratum
feed                             8    6711727.     838966.   15.14
Residual                        62    3435570.      55412.

Total                          107      3.E+07


* MESSAGE: the following units have large residuals.

pig 25          -295.    s.e. 141.

pig 3    period 3             450.    s.e. 178.
pig 4    period 3            541.   s.e. 178.
pig 21   period 3           -526.   s.e. 178.


***** Tables of means *****

Variate: feedint

Grand mean  1093.

     feed        A0        A1        A2         B0        B1        B2
                984.     1228.     1067.      1332.     1260.      687.

                 C0        C1        C2
               1373.     1237.      669.


*** Standard errors of differences of means ***

Table                 feed
rep.                    12
s.e.d.                96.1
```

What has happened is that since information for estimating the means is taken only from the lowest stratum in which the corresponding term in the ANOVA occurs, the within-pig information only is used to produce the difference between each treatment mean and the overall mean. This has the effect of forcing the means for the A, B and C sets of treatments (i.e. the mean of A0, A1, A2, the mean of B0, B1, B2 etc) to be the same. If we compare with the analysis produced if the pig stratum is omitted, the difference is clear.

```
  13  BLOCK period
  14  TREAT feed
  15  ANOVA feedint


15..................................................................

***** Analysis of variance *****

Variate: feedint

Source of variation           d.f.         s.s.        m.s.     v.r.

period stratum                   2     7300968.    3650484.    63.36

period.*Units* stratum
feed                             8       1.E+07    1859787.    32.28
Residual                        97     5588786.      57616.

Total                          107       3.E+07


* MESSAGE: the following units have large residuals.

period 2    *units* 25          -623.    s.e. 227.
period 3    *units* 4            781.    s.e. 227.
period 3    *units* 21          -643.    s.e. 227.


***** Tables of means *****

Variate: feedint

Grand mean  1093.

     feed        A0        A1        A2         B0        B1        B2
               1371.     1616.     1454.      1169.     1097.      524.

                 C0       C1         C2
```

```
             1149.     1013.      445.


*** Standard errors of differences of means ***

Table                    feed
rep.                       12
s.e.d.                   98.0
```

Dropping the pig effect has only slightly affected the residual mean square, implying that the pig effect is small.

Since A0, B0 and C0 do not involve giving any of the feed supplement, they are in fact the same treatment, and this facilitates comparisons of the other treatments. We shall refer to it as ABC0. We really only have seven treatments, and would like to use this in our analysis. However, viewed in this way the design is not balanced, and an analysis using ANOVA is not easily achieved. The REML technique may be used here. The analysis will produce estimates of the different sources of variation in the experiment (pigs, periods, within-pig variation) and the treatment means will be estimated with information being drawn from all strata.

An important issue in specifying the analysis required in REML is whether factors should be considered as fixed or random. This issue is hidden when the experiment is analysed by the ANOVA directive. Block and treatment factors will very often be random and fixed, respectively, if the analysis is performed in REML. The basic idea is that a fixed effect is one you are interested in studying, and for which the essential summary is the table of (estimated) means for each level of the factor. The experiment was performed in order to answer questions about the fixed effects. In the present situation, the feed treatments are a fixed effect.

The random effects correspond to other sources of variation in the experimental material. Usually the levels of a random effect will have been chosen at random from an appropriate population, and the levels of the fixed effects will have been assigned to them. An essential aspect of random effects is that we regard the expected difference between two levels of the effect as zero. In the present case, pigs are a random effect. They have been selected at random from a population of pigs, and assigned at random to the treatment schedules. Often it is obvious whether a factor should be fixed or random, but sometimes it is not so clear. Should period be fixed or random? If the different periods play the role of replicates, then we could consider them as a random effect. In this case the pigs were still growing, and the expected feed intake was not the same in each period, and so periods were considered as a fixed effect. For an unconfounded design, it would make no difference to the estimated treatment effects which effects are regarded as fixed and which random. For partially confounded designs it can make a very important difference.

The following output shows three variations on analysing the feed intake data. In the first, the pig effect is ignored, so that the means are the same as would be obtained by TABULATE. In the second, pig is treated as a random effect, and in the third as a fixed effect.

```
  16  "
 -17   Set up 7 level FACTOR for feed treatments
 -18  "
  19  FACTOR [LEVELS=7; LABELS=!T(ABC0,A1,A2,B1,B2,C1,C2)] feed1
  20  CALCULATE feed1=NEWLEVELS(feed;!(1,2,3,1,4,5,1,6,7))
  21  "
 -22   Analysis with no pig effect
 -23  "
  24  VCOMPONENTS [FIXED=feed1,period]
  25  REML [PRINT=c,s,m] feedint


 25...................................................................


 *** Estimated Components of Variance ***

                                                s.e.
 *units*                        60122.         8545.
```

```
*** Approximate stratum variances ***

                                                 Effective d.f.
*units*                                 60122.3       99.00


* Matrix of coefficients of components for each stratum *

*units*       1.000



*** Table of mean effects for feed1 ***


      feed1         ABCO        A1         A2          B1          B2

                    1229      1616       1454        1097         524

                      C1        C2

                    1013       445

Standard error of differences:    Average          94.85
                                  Maximum          100.1
                                  Minimum          81.73

Average variance of differences:                   9066.


*** Table of mean effects for period ***


     period            1          2          3
                     734       1056       1371

Standard error of differences:      57.79


  26  "
 -27   Analysis with random pig effect
 -28  "
  29  VCOMPONENTS [FIXED=feed1,period] RANDOM=pig
  30  REML [PRINT=c,s,m] feedint

30..............................................................


*** Estimated Components of Variance ***

                                          s.e.
pig                             4190.    6512.
*units*                        55872.    9819.
```

```
*** Approximate stratum variances ***

                                               Effective d.f.
pig                                 68286.1        34.24
*units*                             55871.6        64.76


* Matrix of coefficients of components for each stratum *


pig           2.949       1.000
*units*       0.000       1.000



*** Table of mean effects for feed1 ***


      feed1        ABCO          A1          A2          B1          B2

                   1229        1606        1444        1101         529



Standard error of differences:     Average             93.73
                                   Maximum             99.82
                                   Minimum             80.16

Average variance of differences:                       8860.

*** Table of mean effects for period ***


     period             1           2           3
                      734        1056        1371

Standard error of differences:       55.71


  31  "
 -32   Analysis with fixed pig effect
 -33  "
  34  VCOMPONENTS [FIXED=feed1,period,pig]
  35  REML [PRINT=c,s,m] feedint

35..................................................................


*** Estimated Components of Variance ***

                                           s.e.
*units*                      53681.      9490.



*** Approximate stratum variances ***

                                               Effective d.f.
```

```
*units*                             53680.8        64.00


* Matrix of coefficients of components for each stratum *

*units*       1.000



*** Table of mean effects for feed1 ***


      feed1        ABCO          A1          A2          B1          B2

                   1229        1474        1312        1157         585

                     C1          C2

                   1094         525

Standard error of differences:     Average           117.0
                                   Maximum           133.8
                                   Minimum           94.59

Average variance of differences:                     14059.


*** Table of mean effects for period ***


     period           1           2           3
                    734        1056        1371

Standard error of differences:       54.61



*** Table of mean effects for pig ***


        pig           1           2           3           4           5
                   1182        1108        1170        1436         998

        pig           6           7           8           9          10
                   1488        1125        1328         982        1038

        pig          11          12          13          14          15
                   1314        1178        1125         992         946

        pig          26          27          28          29          30
                    972         848         813        1195        1048

        pig          21          22          23          24          25
                    877         969        1241         892         678

        pig          16          17          18          19          20
                    802         893        1049         981        1009

        pig          31          32          33          34          35
```

```
                      1108         923          1108          925          1183

       pig              36
                      1017

Standard error of differences:      Average             199.6
                                    Maximum             204.3
                                    Minimum             189.2

Average variance of differences:                       39877.
```

The analysis considering pig as a random effect has made only a small modification to the treatment means, when compared with the analysis which ignored pig effects. However, the analysis where the pig effect was considered as fixed is strikingly different. The means have all moved nearer to that of the ABC0 treatment, which has not been affected (because it was observed on all the pigs). To help comparison, the estimates of the means are reproduced in the table below.

Estimated mean feed intake

Treatment	No pig effect	Pig effect random	Pig effect fixed
ABC0	1229	1229	1229
A1	1616	1606	1474
A2	1454	1444	1312
B1	1097	1101	1157
B2	524	528	585
C1	1013	1019	1094
C3	445	450	525
SED	95	94	117

When we omit the pig effect (analysis 1), we assume that the component of variance for pigs is zero and thus give equal weight to the information on treatment effects between pigs and within pigs. The other extreme is to use only the information on treatment effects within pigs, effectively assuming that the component of variance for pigs is infinite. We achieve this by making pig a fixed effect (analysis 3). Both of these analyses ignore the possibility of estimating the pig component of variance, and then using it to determine the weights used to estimate the treatment means. For this to be appropriate, the pigs would require to have been drawn randomly from a population and assigned to their treatment sequences. In this case expected differences between pigs are zero. This assumption is not made in analysis 3. As the assumption was considered appropriate in this experiment, analysis 2 was preferred.

4. Conclusion

This example has illustrated the use of the REML directive to combine the analysis of a series of experiments which when viewed as a whole are not balanced. In unbalanced designs it is intrinsically more difficult to extract information. REML is a powerful technique for handling these situations, but it must be used with care. The issue of fixed versus random effects is one that needs careful thought.

5. References

Box G E P, Hunter W G and Hunter J S (1978) *Statistics for experimenters: an introduction to design, data analysis and model building* Wiley, New York.

Patterson H D and Thompson R (1971) Recovery of inter-block information when block sizes are unequal *Biometrika* **58** 545–554.

Robinson D L (1987) Estimation and use of variance components *The Statistician* **36** 3–14.

A Method of Optimal Categorization of Discrete Variables

Paweł Krajewski
Polish Academy of Sciences
Institute of Plant Genetics
Strzeszyńska 34
POZNAŃ
Poland   60-479

(158)

1. Introduction

Log-linear model fitting is widely used for the analysis of multidimensional contingency tables. If a model is found to be unsatisfactory, the question arises whether the same model can describe the structure of some subtables of the original table. In particular, one can look for a partition of the table into an exclusive and exhaustive set of subtables that would explain the poor fit of the model.

In this paper, a method of optimal partition of the contingency table based on a decomposition of the likelihood-ratio statistic is described. The decomposition, discussed by Gabriel (1966) and Gilula (1985) for two-dimensional tables, has a very clear interpretation in terms of within-group and between-group variability. The interpretation depends, however, on the nature of the variables generating the contingency table. Here, we are particularly interested in the case when partition of the table is defined by groups of categories of one of the variables. Further, we explore the interpretation of the optimal partition when the "partitioning" variable is a random variable. Some consequences of this interpretation are used to construct a method of graphical visualization of the contingency table data. Finally, remarks concerning the generalization of the method for multidimensional tables are given.

2. The Optimality Criterion

Consider a two-dimensional contingency table $T = \{y_{ij}\}$ generated by discrete variables A_1 and A_2 with I and J categories, respectively. Denote the likelihood-ratio statistic for testing the log-linear model M of independence (or homogeneity) by $G^2(M[T])$. It can be shown that for any partition of the set of categories of one of the variables, A_1 or A_2, into mutually exclusive and exhaustive groups, $G^2(M[T])$ decomposes into two parts, which describe the within- and between-group variability, respectively (Gilula 1985). Specifically, if categories of A_1 are subdivided into R groups, G^2 can be written as

$$G^2(M[T]) = \sum_{r=1}^{R} G^2(M[T_r]) + G^2(M[T]|M[T_1,...,T_R]) \tag{1}$$

where T_r denotes the subtable of T corresponding to the rth group of categories of A_1. The first term in (1), denoted by $G^2_W(M)$, is the statistic for testing the simultaneous fit of M on T_r, $r = 1, \ldots, R$; it can be interpreted as a "within-group" component of $G^2(M[T])$. For $r = 1, \ldots, R$ we have

$$G^2(M[T_r]) = \sum_{i=1}^{I} x_{ir} \sum_{j=1}^{J} y_{ij} \log \frac{y_{ij} z_{r+}}{y_{i+} z_{rj}}$$

where

$$z_{rj} = \sum_{i=1}^{I} x_{ir} y_{ij}, \quad z_{r+} = \sum_{i=1}^{I} x_{ir} y_{i+} \quad \text{and} \quad y_{i+} = \sum_{i=1}^{J} y_{ij}$$

with $x_{ir} = 1$, if the ith category is in the rth group, and 0 otherwise. The second part of (1), a conditional statistic denoted by $G^2_B(M)$, can be written as

$$G^2_B(M) = 2 \sum_{r=1}^{R} \sum_{j=1}^{J} z_{ri} \log \frac{z_{rj} y_{++}}{z_{r+} y_{+j}}$$

and is equal to the statistic for testing the fit of M on the $R \times J$ table T_0 constructed from T by summing the observed counts over categories from the same groups.

We define the optimal partition of the categories of A_1 as this partition into R groups, for which $G^2_W(M)$ is minimal. Naturally, the optimal partition maximizes the value of the statistic $G^2_B(M)$. It can be found numerically by searching through all partitions of I categories into R groups, or, for large tables, by two methods described by Siatkowski and Krajewski (1989): successive relocation of categories and minimization of the function $G^2_W(M)$ by the steepest descent algorithm.

3. Interpretation

The meaning of the optimal partition is obvious if the variable A_1 is a factor; that is, if the contingency table comes from the observation of I independent multinomial distributions. In this situation, the statistics $G^2(M[T_r])$ are independent, and the optimal partition of categories of A_1 provides the best grouping with respect to the within-group homogeneity. The meaning of the optimality criterion is similar to that of minimum within-group sum of squares criterion used for clustering multivariate normal populations (Gordon 1982, p 39).

In the case when A_1 is a random variable, and A_2 is a factor, the meaning of the optimal partition is different. In this situation, the statistics $G^2(M[T_r])$ are asymptotically independent (Goodman 1968). The statistic $G^2(M[T])$ describes the "deviance" from the homogeneity of objects corresponding to categories of A_2 with respect to the distribution of A_1. Although the categories of A_1 are defined *a priori*, after the experiment we can ask if we really have to distinguish I categories of A_1 in order to detect the inhomogeneity of objects. Or: how to group the categories of A_1 into R new categories in order to preserve maximum discrimination between objects. It can be seen that the new "categorization" of A_1 we are looking for corresponds to the optimal partition described in Section 1. The amount of information about the inhomogeneity of objects preserved by the new categorization can be measured by the ratio

$$\phi = G^2_B(M)/G^2(M[T])$$

the value of which lies between 0 and 1.

Note that the optimal categorization of a discrete variable can be seen as a reduction of dimensionality, leading from I categories of A_2 to $R < I$ categories. The reduction is done by defining a new (observable) random variable, say $\tilde{A}_1$, which is a linear function of A_1 and has also a multinomial distribution.

4. Graphical Representation

The optimal categorization leads to the possibility of a simple graphical representation of the contingency table data. If $I = 3$, and we reshape our data from counts to proportions, the points representing the categories of A_2 lie on a two-dimensional simplex (Figure 1). This simplex can be projected onto a plane, thus giving a picture of the data convenient for visual inspection (Figure 2). If $I > 3$, this can not be done, but we can use the optimal partition of categories of A_1 into three groups to obtain the graph. The representation we get is optimal in the sense that it preserves maximum discrimination among categories of A_2.

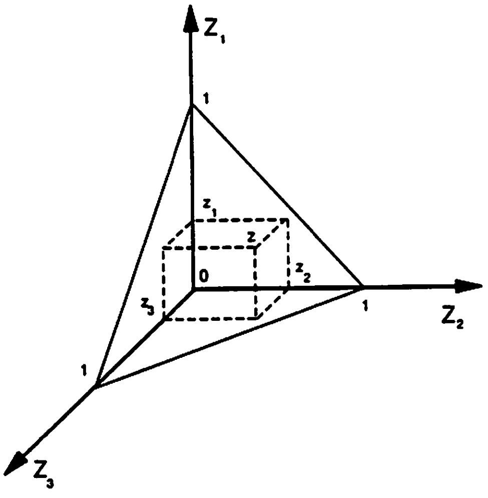

Figure 1. The two-dimensional simplex spanned on unit vectors

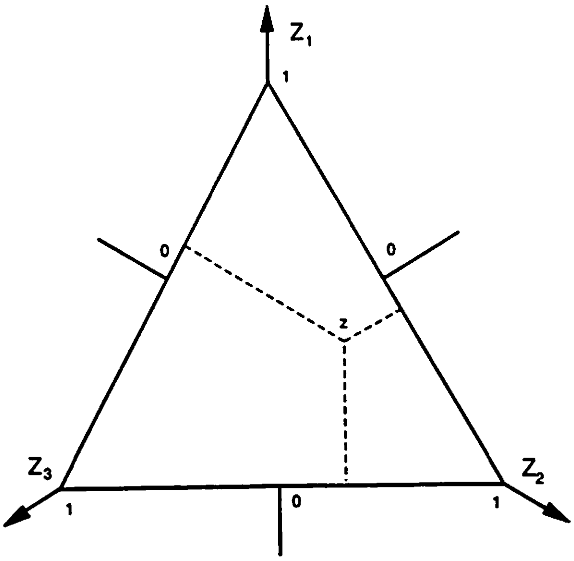

Figure 2. The two-dimensional simplex projected onto a plane

5. Example

Flowers of 12 rose varieties were classified according to their quality into four groups (Table 1). The purpose of the analysis of this contingency table is to detect the differences between varieties (variable A_2) with respect to the quality of flowers (variable A_1).

The value of the G^2 statistic for the homogeneity model is 148.9 with 33 degrees of freedom and is significant. Thus the varieties differ with respect to the quality of flowers, and it is reasonable to ask which grouping of varieties is the best. Using the method of optimal partition into three groups for the variable A_2 we get the following groups of varieties: A = $(1,6,8,12)$; B = $(2,5)$; C = $(3,4,7,9,10,11)$.

Now let us find the categorization of the variable A_1 which preserves maximum information about inhomogeneity of varieties. The best partition of 4 categories of A_1 into 3 groups is: I = (1); II = (2); III = (3, 4), with $G^2_B(M) = 135.9$. Using this categorization of "quality" we preserve $\phi = (135.9/148.9) \times 100 =$ 91.23% of information about inhomogeneity of varieties. The graphical representation of varieties based on this categorization is shown in Figure 3. The positions of the varieties are consistent with their grouping obtained before. It can be seen that groups B and C consist of varieties with the highest and the lowest proportion of flowers in the first quality category, respectively.

Variety (A_2)	Quality of flowers (A_1)			
	1	2	3	4
1	39	42	7	15
2	38	1	1	3
3	3	4	9	4
4	2	22	6	6
5	12	1	1	4
6	19	12	5	3
7	19	19	13	8
8	22	13	2	2
9	12	9	10	13
10	9	15	7	10
11	13	24	6	10
12	12	3	4	1

Table 1. Flowers of 12 rose varieties classified according to their quality into four categories

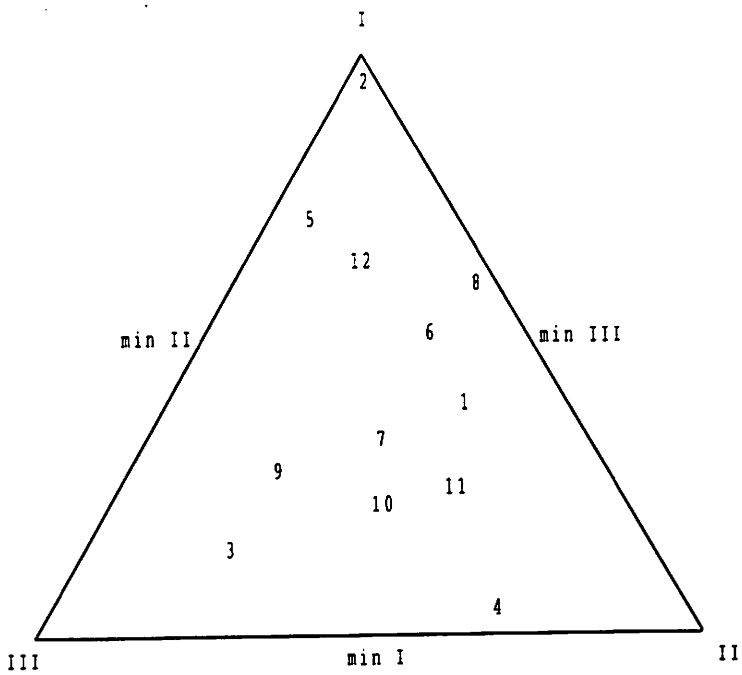

Figure 3. Graphical representation of varieties in the triangular coordinate system corresponding to optimal categorization (1), (2), (3, 4) of the variable "quality of flowers" (obtained using Genstat procedure DTRIA)

6. Generalization

In the general case, we consider a multidimensional contingency table T generated by S discrete variables. If we take any of these variables, say A_q, and divide the set of its categories into R subgroups, the decomposition (1) holds for any hierarchical log-linear model M of the table T. Thus, we can look for the partition of the categories of A_q optimal with respect to the model M. Such generalization does not introduce special theoretical problems, although it turns out that the search for optimal partition and proper interpretation of the results are possible only for a subclass of log-linear models. Some facts relevant to this problem are summarized below. Full description of the method of optimal partition of multidimensional contingency tables is given by Krajewski (1989).

For a two-dimensional table, the statistic $G_B^2(M)$ can be calculated from the table T_0 defined in Section 2 and can be interpreted directly as a between group statistic. This is not true for the general case. It can be shown that this property holds only for the so called q-collapsible models characterized by the generating class consisting of two elements, one of which does not contain q. The homogeneity model of a two-dimensional table is both 1- and 2-collapsible; for a three-dimensional table, models characterized by generating classes $\{12, 23\}$, $\{13, 23\}$ and $\{1, 23\}$ are 1-collapsible models. Because the $G_B^2(M)$ statistic for the optimal categorization of a variable should be interpretable in terms of inhomogeneity of factor levels, the grouping of the categories of the random variable A_q should be used only for q-collapsible models.

For some models, the value of $G_W^2(M)$ is the same for any partition of categories of A_q, which makes the search for optimal partition meaningless. It can be shown that this happens if the model is q-decomposable, that is, if all elements of its generating class contain q (example: model $\{12, 13\}$ of a three-dimensional table).

As far as the graphical representation of data is concerned, for the case of a multidimensional table, we can obtain a graph which represents combinations of the categories in the simplex spanned by the categories of the random variable. If a graph is made in the triangular coordinate system, corresponding to the optimal partition of the categories of the random variable, its interpretation depends on the model for which optimal grouping was found.

7. References

Gabriel K R (1966) Simultaneous test procedures for multiple comparisons on categorical data *J. Amer. Statist. Assoc.* 61 1081–1096.

Gilula Z (1985) On the analysis of heterogeneity among populations *J. R. Statist. Soc.* 47 76–83.

Goodman L A (1968) The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries *J. Amer. Statist. Assoc.* 63 1091–1131.

Gordon A D (1982) *Classification* Chapman and Hall, London.

Krajewski P (1989) *Statistical analysis of multidimensional contingency tables (In Polish)* Institute of Plant Genetics, Polish Academy of Sciences.

Siatkowski I and Krajewski P (1989) Grouping multinomial objects according to minimal within-group variability *Statistica Applicata* 1 33–43.

8. Procedure

Note: The Genstat procedure presented below works only for two-dimensional tables. It draws a picture of factor categories in the triangular coordinate system corresponding to given categorization of the random variable. The procedure does not look for the optimal categorization (relevant Fortran 77 procedures are available from the author on request).

```
PROCEDURE 'DTRIA'
OPTION NAME='TITLE','DIST'; MODE=t,v; DEFAULT=' ',0.02
PARAMETER NAME='OBJ','CAT','DATA','GROUPS','MINIMA','GTOT','GBET'; MODE=p

"The procedure analyses contingency table generated by two variables,
one of which is a random variable and the other is of factor type.
It draws the picture of the points representing factor categories
in a triangular coordinate system corresponding to given categorization
of the random variable.

Options:
TITLE  : string;  the title of the graph (default=*),
DIST   : number;  the smallest distance between a point in the graph
and its edge (default=0.02),

Input parameters:
OBJ    : scalar;  the number of factor categories,
CAT    : scalar;  the number of categories of random variable,
DATA   : variate; data from the OBJxCAT contingency table
(row after row),
GROUPS : variate; gives the new categorization (group numbers),

Output parameters:
MINIMA : variate; minimal values for axes obtained,
GTOT   : scalar;  the value of G2 statistic for original categorization,
GBET   : scalar;  the value of G2 statistic for given categorization"

"Prepare factors classifying the table"
CALCULATE n=OBJ*CAT
FACTOR [NVALUES=n; LEVELS=OBJ] a
& [NVALUES=n; LEVELS=CAT] b
GENERATE a,b

"Fit the model for the original categorization"
MODEL [DISTRIBUTION=poisson] DATA
FIT [PRINT=*] a+b
RKEEP DATA; DEVIANCE=GTOT

"Convert the data to a table"
TABULATE [CLASS=b,a] DATA; TOTALS=tablica

"Calculate the table corresponding to the new categorization"
CALCULATE fa=OBJ*3
FACTOR [NVALUES=fa; LEVELS=3] bb
& [LEVELS=OBJ] aa
GENERATE bb
```

```
COMBINE [OLDSTRUCTURE=tablica; NEWSTRUCTURE=tab] \
  OLDDIMENSION=b; NEWDIMENSION=bb; \
  OLDPOSITIONS=!(1...CAT); NEWPOSITIONS=GROUPS
VARIATE [NVALUES=OBJ] y[1],y[2],y[3]
EQUATE OLD=tab; NEW=y

"Fit the model for the new categorization"
VARIATE [NVALUES=fa] yy
EQUATE OLD=y; NEW=yy
GENERATE bb,aa
MODEL [DISTRIBUTION=poisson] yy
FIT [PRINT=*] aa+bb
RKEEP yy; DEVIANCE=GBET

"Convert the data to proportions"
CALCULATE ysum=y[1]+y[2]+y[3]
FOR i=1,2,3
  CALCULATE y[i]=y[i]/ysum
ENDFOR

"Find the part of the full simplex containing data points"
SCALAR h
CALCULATE h=1
FOR i=1,2,3
  CALCULATE ss=min(y[i])-DIST
  IF  ss<0.0
    CALCULATE ss=0.0
  ENDIF
  CALCULATE h=h-ss
  CALCULATE MINIMA$[i]=ss
ENDFOR

"Calculate points coordinates in the rectangular system"
FOR i=1,2
  CALCULATE y[i]=(y[i]-MINIMA$[i])/h
ENDFOR
CALCULATE y[2]=1.1547*(y[2]+y[1]/2)

"Set the points for vertices of the triangle"
VARIATE [NVALUES=4] xtr; VALUES=!(0,1.155,0.577,0)
VARIATE [NVALUES=4] ytr; VALUES=!(0,0,1,0)

"Generate the numbers of points"
FACTOR [NVALUES=OBJ; LEVELS=OBJ] num
GENERATE num

"Generate the description of the vertices"
TEXT [NVALUES=3] des; VALUES=!t(I,II,III)
VARIATE [NVALUES=3] xdes; VALUES=!(0.572,1.18,-0.05)
VARIATE [NVALUES=3] ydes; VALUES=!(1.05,-0.05,-0.05)

"Generate the description of the axes"
TEXT [NVALUES=4] minval; VALUES=!T('min I','min II','min III',' ')
VARIATE [NVALUES=4] xxmin; VALUES=!(0.54,0.155,0.88,0)
VARIATE [NVALUES=4] yymin; VALUES=!(-0.05,0.5,0.5,0)

"Set the graph parameters"
PEN 1,2,3,4; COLOUR=1,1,1,1; LINE=1,0,0,0; METHOD=line,point,point,point;\
  JOIN=given; SYMBOLS=0,num,des,minval
FRAME 1; 0; 1; 0; 1
AXES 1; YLOWER=-0.10; YUPPER=1.3; XLOWER=-0.10; XUPPER=1.3; STYLE=none

"Draw the picture"
DGRAPH [TITLE=TITLE; WINDOW=1; KEYWINDOW=0] ytr,y[1],ydes,yymin;\
  xtr,y[2],xdes,xxmin; PEN=1,2,3,4

ENDPROCEDURE
```

9. Appendix

The main program

```
"An example program using the procedure DTRIA"

SCALAR na,nb,dist,gtot,gbet,phi
VARIATE [NVALUES=3] minima
READ na,nb,dist
12 4 0.03:
VARIATE [NVALUES=nb] groups
READ groups
1 2 3 3:
CALCULATE npod=na*nb
VARIATE [NVALUES=npod]  x
READ x
39 42 7 15 38 1 1 3 3 4 9 4 2 22 6 6
12 1 1 4 19 12 5 3 19 19 13 8 22 13 2 2
12 9 10 13 9 15 7 10 13 24 6 10 12 3 4 1:

DTRIA [TITLE=' '; DIST=dist] OBJ=na; CAT=nb; DATA=x; GROUPS=groups;\
  MINIMA=minima; GTOT=gtot; GBET=gbet

FACTOR [NVALUES=nb; LEVELS=nb] categ
TEXT [NVALUES=3] des; VALUES=!t(I,II,III)
GENERATE categ
PRINT 'Original categories and their grouping :'
PRINT [IPRINT=*] categ,groups
PRINT 'Minimal values for axes :'
PRINT [IPRINT=*] des,minima
PRINT 'G2 statistic values :'
PRINT [IPRINT=*] 'Original categorization : ',gtot; FIELD=26,7; DECIMALS=3
PRINT [IPRINT=*] 'New categorization      : ',gbet; FIELD=26,7; DECIMALS=3
CALCULATE phi=(gbet/gtot)*100
PRINT [IPRINT=*] 'Information preserved   : ',phi,'%'; FIELD=26,7,2; \
  DECIMALS=2

STOP
```

Edited output (text output only, see Figure 3 for graphics output)

```
Original categories and their grouping :

1       1.000
2       2.000
3       3.000
4       3.000

Minimal values for axes :

I     0.02556
II     0.00000
III     0.06302

G2 statistic values :

Original categorization :  148.947
New categorization      :  135.879
Information preserved   :   91.23 %
```

Handling Hierarchical Data

B M Church
Statistics Department
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
HARPENDEN
Herts AL5 2JQ

1. Introduction

Data from sample surveys or other observational data are often hierarchical. For example, information may be available about household characteristics for a sample of households (level 1) with additional information about each individual in the household (level 2). It may be convenient or compact to store these data in separate computer-files linked only by common reference numbers. Thus, the first file would contain one record for each unit at level 1 (that is, only one record for each reference number) while the second file would contain a variable number of records, from 0 upwards, having the same reference number as a record at level 1. Usually both files are in reference-number order but this is not assumed. Frequently one needs to process information from both levels together, and this may be achieved in two ways:

1. Required information from the top level, level 1, may be 'pushed down' to level 2. This means that information from each record at level 1, is replicated once for each record with the same reference number, at level 2.

2. Summary information from level 2 may be 'pushed up' to level 1.

In either case, the key is to define a factor with levels corresponding to values of the reference number.

2. Notation

The top level of hierarchy is referred to as level 1 and the lower as level 2; data files corresponding to these levels are referred to below as i1 and i2. The data are assumed to be in backing-store files, which is sensible for large data sets, as only the required variates are accessed. The procedures outlined below however, are equally applicable to character files.

Variates recorded, or in use at levels 1 and 2 are **x1[]** and **x2[]**. Reference numbers ref1, ref2 may be read or constructed from **x1[]**, **x2[]** respectively; they are referred to below as if read explicitly by name. Variates **x1[]**, or functions of these variates, are referred to as **v2[]** when pushed down to level 2; similarly summaries of **x2[]** are referred to as **v1[]** when pushed up to level 1.

3. Pushing Down

The following instructions push down selected variates **x1[]** from level 1 for direct use in analysis (tabulation etc) with variates at level 2, or to be stored as **v2[]** on a secondary file or subfile for subsequent use.

```
RETRIEVE [CHANNEL=i1] ref1,x1[]
SORT [INDEX=ref1; GROUP=reff; LEVELS=reflev]
RETRIEVE [CHANNEL=i2] ref2,x2[]
FACTOR [NVALUES=ref2; MODIFY=yes] reff
CALCULATE reff = ref2
& v2[] = NEWLEVELS(reff; x1[])
```

Values of **v2[]** are set missing (and warning **VA3** given) for values of ref2 which do not occur in ref1. These instructions are appropriate only if values from level 1 are not needed for combined analysis when there are no corresponding data at level 2; this is usually so: see below.

4. Pushing Up

The first five lines of code remain as above and are followed by:

```
TABULATE [CLASS=reff] x2[]; NOBS=tc[]; MEANS=tm[]; TOTALS=tt[]
```

where counts, means and totals may be needed for different subsets of x2[].

```
VARIATE v1[]; tc[],tm[],tt[]
```

where there is one v1[] for each tc[], tm[], tt[]. Then x1[] and v1[] may be used directly for analysis (tabulation etc), or written to a secondary file for subsequent use.

This procedure produces zeros for counts and missing values for means in the v1[] for reference numbers which occur only at level 1. Data corresponding to reference numbers occurring only at level 2 are treated as invalid and are omitted.

One may sometimes need to push up several variates corresponding to one variate x2[i]. For example, if crop type and area are recorded for individual fields at level 2, one may wish to push up total areas on the farm under each crop type (e.g. for farm-type characterisation). This is achieved by

```
TABULATE [CLASS=reff,cropf] area; TOTALS=tt
VARIATE [NVALUES=reflev] v1[]
EQUATE OLD=tt; NEW=v1
```

where cropf is a factor for crop and there is one v1[] for each level of cropf.

5. Mismatched Data

For most purposes, even when there is not a good match between levels 1 and 2, the above procedures suffice if primary data files are retained for analyses requiring data at only one level. However, sometimes — for example, when levels 1 and 2 comprise similar variables recorded on two occasions rather than a genuine hierarchy — it may be desirable to retain unmatched data in a secondary file. As an example, channels i1 and i2 might both contain reference numbers and descriptive data (region, cropping pattern, and so on) for soil-sampled fields, with analytical determinations for the soil samples at laboratories A and B respectively. In a comparison between laboratories, most information would be from matched samples but some might be recovered from fields for which determinations were made in only one laboratory.
A suitable secondary file may be constructed as follows:

```
RETRIEVE [CHANNEL=i1] ref1,x1[]
RETRIEVE [CHANNEL=i2] ref2,x2[]
VARIATE ref; VALUES=!(#ref1,#ref2)
SORT [INDEX=ref; GROUP=reff; LEVELS=reflev]

FACTOR [NVALUES=ref1; LEVELS=reflev] reff
CALCULATE reff = ref1
TABULATE [CLASS=reff] x1[]; MEANS=t1[]
VARIATE v1[]; VALUES=t1[]
DELETE [REDEFINE=yes] reff,t1[]

FACTOR [NVALUES=ref2; LEVELS=reflev] reff
CALCULATE reff = ref2
TABULATE [CLASS=reff] x2[]; MEANS=t2[]
VARIATE v2[]; VALUES=t2[]
DELETE [REDEFINE=yes] reff,t2[]

STORE [CHANNEL=i3] v1[],v2[]
```

6. More Complex Hierarchical and Other Datasets

These may be tackled by writing your own Fortran subroutine OWNTAB (see Genstat 5 Release 2 Reference Manual Supplement, Section 5.4.4).

Constructing a System within Genstat

J A Nelder
Department of Mathematics
Imperial College
180 Queen's Gate
LONDON
United Kingdom SW7 2BZ

1. Introduction

This note describes the properties of a system I have built in Genstat; it is called the K system and it has been constructed to provide an environment for doing intensive interactive work with GLMs. Full details for running the system are given in a NAG Technical Report, and this should be consulted by anyone wanting to use the system. This note concentrates on some of the general ideas behind it and how they were implemented.

My over-riding aim in developing the K system has been to reduce the amount of typing by the user. The less typing required, the less chance of making mistakes and the more quickly one can work. I have been an intensive user of GLIM, and it has some features that are valuable for my particular mode of working. The first, and most important, is that system vectors such as %fv for the fitted values are set automatically after a fit; thus no `RKEEP` instructions need follow a `FIT`. This saves a lot of typing, and avoids having to make up names for the identifiers required by `RKEEP`.

The second feature is the reduction in the amount of output. If you are trying many models during an exploratory phase then the deviance and d.f., plus the change from the previous fit, when relevant, are usually all that are needed. When a model looks promising, then one needs estimates of parameters, a listing of residuals, the variance-covariance matrix, etc. With minimal output the screen can hold details of the fits of several models without any need to scroll back; reduction of typing again.

Both GLIM 3.77 macros and Genstat procedures have parameters; in GLIM these are always of the form %1, %2..., while in Genstat they are identifiers with capital letters. However, a major difference between a GLIM 3.77 macro and a Genstat procedure is that in the GLIM macro all the other identifiers are global, whereas in a Genstat procedure they are all local. Both conventions have advantages and disadvantages. Having global identifiers means that system vectors like %fv are automatically available and do not have to be passed as parameters; conversely there is a danger that when several macros appear in a GLIM program they may accidently use the same identifiers to mean different things. Having local identifiers means that what goes on in a procedure is completely insulated from the outside world, so that cross-contamination becomes impossible; however, this also means that all external information has to be passed in via parameter lists, and parameter lists can take a lot of typing. Fortunately, this last statement turns out not to be entirely true in Genstat; there is a way of defining and using what are equivalent to COMMON blocks in Fortran, and with these we can define sets of global identifiers for sets of procedures.

2. The Hidden-option Trick

Suppose we want the identifiers `a`, `b`, and `c` to become common to a set of procedures. We first define a pointer p, say, whose values are `a`, `b`, and `c`. This pointer must be defined outside any procedure, so that it continues to exist on exit from any procedure that requires it. In any procedure that needs p, we define a hidden option `P`, coming after any that we may need to provide for the user, and set its default value to p. Thus suppose that procedure `PROC` also has an 'open' option `O`, known to the user. Its definition would begin

```
PROCEDURE 'PROC'
OPTION 'O','P'; DEFAULT= *,p
```

(This definition assumes no default for `O`.) Within the procedure the identifier b, for example, may be referred to as P[2], but it makes writing the procedure much easier if the body of the procedure begins with a statement of the form

```
DUMMY a,b,c; VALUE=P[]
```

Now a, b, and c can be referred to by their 'real' names.

One of the 'common blocks' in the K system is called glmmdl, and it holds items required to define a GLM. The definition, which is held in a file called kinit, has the following form:

```
POINTER glmmdl; VALUES=!p(%err,%lin,%exp,%sca,%pw,%os,%yv,%bd,%ter,%nu,%dpw,%dos,\
                         yvid,bdid,osid,pwid)

FORMULA %ter
POINTER yvid,bdid,osid,pwid
SCALAR  %exp,%sca,%nu
TEXT    %err,%lin
VARIATE %pw,%os,%yv,%bd,%dpw,%dos
```

Many of the names will be familiar to GLIM users. The pointers are used for internal housekeeping, while %ter is used to hold the maximal model as defined in TERMS, something that is not required in GLIM. The two identifiers %dpw and %dos are not GLIM names; they have been invented to hold the default values of the prior weight and offset variates, respectively 1 and 0. Note also that %err and %lin are here of type text, rather than coded integers as in GLIM. Any procedure wishing to make use of the elements of glmmdl must include in its option statement the setting

```
OPTION 'GLMMDL'; DEFAULT=glmmdl
```

and, if the names of the elements are to be used, a DUMMY statement of the type illustrated above.

Other 'common blocks' in the K system are (1) glmrk, which holds the elements relevant to a GLM that can be saved in RKEEP; (2) mcproc, which holds structures relevant to model-checking, for example the Cook's statistics, and (3) glmown, which holds the components required to define an OWN model in the GLIM style.

3. Defaults for Options and Parameters

Typing can be reduced when options and parameters in a procedure call are commonly the same. For example, in the model-checking procedure called npl_, which produces Normal plots, the full specification is

```
npl_[<type>]  <variate1>; <variate2>
```

where type is h or f (for half- or full-Normal plots), variate1 is the variate to be plotted, and variate2 is a weight vector whose zeros indicate units to be excluded. I use overwhelmingly the setting h for the option, the deviance residual (available in %res in glmrk) for the first parameter, and the prior weight (available in %pw) for the second parameter. With these default settings, and provision of the necessary common blocks, the call

```
npl_[h] %res; %pw
```

becomes

```
npl_
```

and the number of characters typed falls from 15 to 4. The provision of intelligently chosen defaults is, I believe, a vital part of a good interactive system. Note that the presence of the common blocks is a vital part of this provision.

4. Reducing Output

Output is reduced in the K system in two ways. In the first, a parallel procedure is set up to a Genstat command, the standard output being suppressed, and replaced by the abbreviated form. Thus corresponding to FIT we have kfit; kfit suppresses the output from FIT and replaces it by the GLIM form, involving just the deviance and d.f., with changes from the last fit, if appropriate. However, kfit does more than this, because it also calls RKEEP to save the structures accessible in the pointer glmrk. The two most important options in FIT — CONSTANT, which defines the treatment of the intercept, and FACTORIAL, which controls the maximum order of terms to be fitted in a factorial linear predictor — can be passed through kfit. They have the same default values as in Genstat. If further options are required it is easy to amend the code accordingly. The procedures kadd, kdrop, and kswit replace ADD, DROP, and SWITCH respectively.

The second way of reducing output is to change the options in a Genstat directive appropriately. A good example is PRINT, which for my purposes produces too much blank space and has a default setting of values in parallel, when values printed serially are much more compact. The K procedure kpr deals with this by printing values serially, suppressing the identifier names, omitting blank lines, and using an orientation across the page. In this way the command

```
 PRINT [SERIAL=yes; IPRINT=*; SQUASH=yes; ORIENTATION=across] a,b,c
```

becomes

```
 kpr a,b,c
```

Similarly I find that when using DELETE I always want the option REDEFINE set to yes, rather than to its default value of no. The K procedure kdel does this.

5. An Example

The K system has sets of procedures for defining models, for fitting them, for inspecting the output, and for checking the consistency of the fit, together with a few general procedures like kdel described above. It is entirely written in Genstat. The file kmake contains commands for constructing a procedure backing-store file, while ksu sets up the system subsequently (and much more rapidly). The following example will give an idea of the syntax. It shows the input only and is commented for purposes of explanation.

```
                    "set up artificial data set "

kdel y,x,xx,a,w           " redefine variables in case already in use "
VARIATE[VALUES=1...50] x
CALCULATE y=10+0.01*x*x+URAND(23;50) & xx=SQRT(x)
FACTOR[LEVELS=2;VALUES=25(1,2)] a

               " use some of the GLM procedures in the K system "

kun 50                    " set number of units and initialise "
yvar y                    " define response variable "
kterm a*(x+xx)            " define maximal model "
kfit x+xx                 " fit x+xx with minimal printing, saving output via rkeep "
kadd a                    " add factor a to linear predictor, etc. "
kdrop x                   " drop term x, etc. "
kswit a,x                 " switch a & x, etc. "
CALCULATE w=(y>10)
wei w                     " set up prior weight omitting units with y<=10 "
kadd                      " re-fit model with same linear predictor "
err p                     " change error from (default) Normal to Poisson "
kadd                      " re-fit model with same linear predictor "
```

```
              "use some of the model-checking procedures "

sumc                       " set up for model-checking procedures "
drp_                       " plot dev. resid. v. function of fitted values "
npl_                       " half-Normal plot of dev. residuals "
npl_[f]                    " full Normal plot of dev. residuals "
CALCULATE x2=x*x
avp_ x2                    " make added-variable plot for x2 "
```