NAG®

GENSTAT

Newsletter

Issue No. 24

Editors

P W Lane
Rothamsted Experimental Station
HARPENDEN
Hertfordshire
United Kingdom AL5 2JQ

K I Trinder
NAG Limited
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom OX2 8DR

Printed and produced by the Numerical Algorithms Group

ISSN 0269-0764

Genstat Newsletter
Issue No. 24

NP1976

October 1989

Contents

Published Twice Yearly by
Rothamsted Experimental Station Statistics Department
and the Numerical Algorithms Group Ltd

Editorial

This issue has been a little delayed because of a shortage of articles. However, articles are now arriving based on talks given at the Genstat Conference in Edinburgh. Three of these have been included in this issue, together with the report on the Conference. We hope that the next issue will appear on schedule with more of these articles; it will also contain information on the new facilities in Release 2.

We are experimenting this issue with a new layout for news items. These will now appear not in the editorial, as in previous issues, but in a News section – hence the shortness of this editorial. So look at that section for the status of implementations of Release 1, a progress report on Release 2, information about the new Procedure Library Manual, and notice of the next one-day Genstat conference.

Two articles in this issue describe experience with the PC implementation of Genstat, and should be of particular interest to PC users of Genstat – who constitute a large and growing proportion of all users of the system. There is also a further article in the series on language features. Finally, the article on use of Genstat in graduate projects at the University of Cambridge should be of particular interest to readers who use Genstat in higher education.

Genstat 5 Introductory Course

Please note that NAG are in the final stages of arranging the next Introductory Course, to be presented in conjunction with staff of Rothamsted Experimental Station. It is expected that this will take place in mid-June or mid-July of 1990 and will be in the Bristol area. For further information, please contact Keith Trinder at NAG Ltd, Wilkinson House, Jordan Hill Road, Oxford, United Kingdom OX2 8DR; tel: 0865 511245 (or +44 865 511245); fax: 0865 310139 (or +44 865 310139).

Genstat News

New Implementations

Several new implementations of Genstat 5 have become available since the last Newsletter. Perhaps the most noteworthy of these is a version for IBM compatible PCs with the 80386 (or 80486) processor. The PC must also have 2 Mb of memory, 10 Mb of free disk space and PC-DOS or MS-DOS 3.1 or higher. A mathematical co-processor is not essential but is strongly recommended. This version is in addition to the one already available for IBM compatible PCs with any processor, although it is much faster and is able to handle much larger problems with a default data space of 524,288 items. It has the full range of Genstat facilities including graphics, extendability with Fortran and the latest Procedure Library, (see below). These two versions have the same price structure with licences being charged at a single one-time fee depending on the level of licence and the optional support service costing a corresponding annual fee.

Genstat 5 is also now available for the IBM MVS, IBM CMS, IBM 6150 AIX and Norsk Data 5000 Sintran III systems. Of these, the IBM MVS implementation has the graphics and Fortran extension facilities. This is also the first version to use double precision throughout for storage of data items, as well as calculations. In other versions, most storage of data items is in single precision and calculations are double precision when calculation errors are likely to accumulate. Sites with Genstat on Prime systems have also been sent Release 1.3 to replace Release 1.2, allowing them to have use of the graphics facility.

Procedure Library Release 1.3[2]

The latest Procedure Library containing 66 documented procedures has recently been sent to sites with any implementation of Genstat 5 Release 1.3. The slightly curious release numbering system reflects the fact that the library can only be used with Release 1.3 of Genstat 5 and that this is the second such library. (There have been two other libraries: Releases 1.3[1] with 51 procedures and Release 1.2[1] with just seven procedures.) The new library includes procedures for ante-dependence analysis of repeated measures data, Box-Jenkins ARIMA model fitting to time series data, convex hull fitting and for producing scree diagrams and Box and Whisker plots.

The new library is accompanied by a printed manual for the first time and copies of the manual can be obtained from NAG at £5.00 each. Alternatively, the `LIBMANUAL` procedure is still available allowing users to produce the manual in a text file form.

One-day Conference

The next Genstat one-day conference will be about 'Interactive Statistical Modelling'. It will be held on Thursday, 26 April 1990, at the Conference Centre, Rothamsted, Harpenden, Herts. There will be presentations about the new interactive features of Release 2, particularly the introduction of interactive graphics and the menu system. Other talks will concentrate on the ways in which specific areas of statistical analysis can be approached interactively. Registration forms will soon be sent out to all sites, but write direct to the Statistics Department at Rothamsted if you would like one sent to you personally.

Progress on Release 2

Since the Genstat Conference, the prototype version of Release 2 – known as Release 2.0 – has been tested on VAX VMS systems first at Rothamsted and later at all institutes in the Agricultural and Food Research Council. At the same time, the Fortran source code has undergone portability checks using NAG's Toolpack system, to eliminate many of the potential difficulties faced by implementors on other computer systems. Release 2.1 will be formed at Rothamsted early in 1990, and passed to NAG for certification. The base version should be sent to implementors in the Spring, and the first implementations should be distributed to sites soon after.

Sixth Genstat Conference, Edinburgh, 11-15 September 1989

J Maindonald
Applied Mathematics Division
DSIR
Auckland
New Zealand

To judge from the papers and presentations at this conference, Genstat 5 is now firmly established. Improvement and development is proceeding apace. There is rapid progress, also, in the writing of Genstat 5 procedures, providing abilities beyond what is immediately available in Genstat itself. This has implications for future development of Genstat. In many areas of work, particularly those that are advancing rapidly, it is appropriate to place the emphasis on providing abilities on which procedure writers can build.

Release 2 was demonstrated on Sun workstations at the conference. This includes new facilities for multi-dimensional scaling, residual maximum likelihood (REML), monotone regression, interactive graphics, and the directive QUESTION for use in constructing menu systems. There are also a host of more minor improvements, such as the ability to manipulate formulae, save contrasts after ANOVA statements, and tabulate hierarchical surveys. An 80386-based PC version of Release 1.3 excited considerable interest.

What of analyses that cannot easily be handled using procedures, that seem to require the writing of separate code in Fortran or another high-level language? This is how REML started; it was first made available as a separate program, with a primitive user interface, before incorporation into Genstat. Will this remain the pattern in future? Or is it now sufficiently easy to add new routines and their associated directives into Genstat that programs of this type will in future be developed *ab initio* as part of Genstat?

Further improvements in the user interface will be required if Genstat is to be a really attractive environment into which to add new routines. An obvious gap is that there is no facility for recalling, editing and re-executing statements previous to the latest statement, although some operating systems do provide this.

The new QUESTION directive is a first step in making Genstat more attractive to non-specialist users. It assists the writing of procedures that offer guidance in handling Genstat analyses. Peter Lane demonstrated a conversational interface, based around the QUESTION directive, for users wanting to do regression with a single explanatory variate.

Another possibility is to write a 'front end' program, separate from Genstat, that will guide users who want to undertake one of a limited and well-defined class of analyses. James Roger demonstrated Datachain, based around an attractive pull-down menu system and providing extensive assistance in the design and analysis of field experiments. This extends to facilitating data entry and writing a file that contains the main part of the Genstat code required for handling the analysis. If Datachain lives up to its promise, it will surely boost sales of Genstat.

Most of us have had the experience of trying to retrace the steps involved in a series of analyses performed some months earlier, of which details have been forgotten. Pete Digby had advice on disciplines for documenting what has been done, easing the path for anyone who in future may want to retrace the same steps.

There were several talks on using Genstat in teaching statistics. Chris Glasbey described how, since October 1988, the Scottish Agricultural Statistics Service has given one of several two-day modules to a total of 240 scientists. 'Teaching and consultancy', we were told, 'are complementary for today's statisticians'. Two further talks, by Lars Thomsen and Christine Donnelly, discussed the use of Genstat in undergraduate university statistics courses in Denmark and Australia. Other overseas experiences were described by Carlos Gonzalez, who gave an account of the use of Genstat at CIMMYT in Mexico.

Two talks reviewed an area of statistical application. Les Underhill provided a lucid commentary on the folklore of non-metric scaling. Robin Thompson reviewed the residual maximum likelihood (REML) approach to estimation of variance components. He mused on possibilities for an equivalent to REML within the generalised linear model framework, to be known as a GREMLIN analysis. Sue Welham explained how REML has been implemented in Release 2.

There was a wide range of talks about the theory and application of new and established statistical techniques. Many concerned regression analysis, particularly generalized linear models, with several contributions from Dutch statisticians. Others focused on multivariate techniques, the analysis of designed experiments, and on Fourier analysis. It is hoped that the content of many of these talks will be published in this issue and future issues of the Genstat Newsletter.

The standard of presentation was high. My one grouse is that many of the overhead projector acetates had lettering that was too small. In a few cases, it was impossibly small. It is just possible that I will bring binoculars to the next conference.

I was surprised that there was no session devoted to getting user views. This had been a feature of the Australasian Genstat workshop. Perhaps the Genstat developers had been sated with user views at the Australasian workshop.

Wednesday afternoon gave a choice from two outings, one to the Glenkinchie distillery, and the other to Roslin chapel followed by a visit to a butterfly farm. There are no prizes for guessing which was the more popular. The manager of the butterfly farm showed a glow of recognition when Genstat was mentioned – he had met it while attending a course at Lincoln College in New Zealand.

The Thursday evening dinner at Dalhousie court claimed to transport us back in time to 1623, when spoons had not been invented, and forks ('you can stab four times with one push') were a novelty. Visitors from a number of other places, including Canada and the Land of US, were welcomed in their native tongues. New Zealand missed out – in 1623 even the Scots knew nothing of it. Roger Payne, as presiding Lord Genstat, missed a PR opportunity by passing up the invitation to be Guest of Honour for the evening. At the end of the meal our waitresses transformed themselves into a choir, rounding off the evening with pleasant singing.

Thanks are due to Jackie Muscott and the Scottish Agricultural Statistics Service for their excellent local organisation.

Modelling the Variance in Regression

M S Ridout
Institute of Horticultural Research
East Malling
Maidstone
Kent
United Kingdom *ME19 6BJ*

1. Introduction

In ordinary linear models, the response variate is assumed to have constant variance. Generalized linear models (GLMs) are more flexible, assuming instead that the variance is proportional to a known function of the mean (the variance function). But although this formulation is sufficiently general for many practical applications, it remains restrictive: partly because it assumes that the variance function is known exactly, but also because it assumes that any variance heterogeneity is entirely attributable to changes in the mean response. This paper includes an example in which the variance heterogeneity is not of this form.

The paper describes a model that allows the variance structure of the data to be modelled more flexibly than in a standard linear model. We discuss various methods of estimating the parameters of the model and show how the methods can easily be implemented using the existing regression facilities in Genstat.

2. The Model

Let $y_1, y_2, \ldots, y_n$ be independent observations with means μ_i and variances σ_i^2. For the means we assume an ordinary linear model

$$\mu_i = x_i^{\mathrm{T}} \beta \qquad (1)$$

where x_i is a vector of p covariates relating to the ith observation and β is a vector of p parameters.

We now suppose that the variance, or dispersion parameter, σ_i^2 also varies between observations, and that this variation can be modelled using a log-linear model

$$\log(\sigma_i^2) = z_i^{\mathrm{T}} \gamma \qquad (2)$$

where z_i is a vector of q covariates and γ is a vector of q parameters. We call this the 'dispersion submodel' to distinguish it from the 'mean submodel' defined by equation (1). The covariates z_i may or may not coincide, either partially or completely with the covariates x_i. The only *a priori* restriction is that the variance should be functionally independent of the mean.

3. Parameter Estimation

We now discuss several methods of estimating the parameters β and γ, and outline the implementation of these methods in Genstat.

3.1. Maximum-likelihood Estimation Assuming Normality

If the data are assumed to be Normally distributed then the log-likelihood is

$$\mathrm{loglik}(y;\ \mu_i, \sigma_i^2) = -\frac{1}{2}\left[(y_i - \mu_i)^2 / \sigma_i^2 + \log(\sigma_i^2) + \log(2\pi)\right]$$

where μ_i depends on the parameters β, and σ_i^2 depends on the parameters γ. At first sight, maximum-likelihood estimation using Genstat appears to require the general optimization facilities provided by the FITNONLINEAR command. However, it can be

shown [1], [11] that the maximum-likelihood estimates of β and γ may alternatively be obtained by the following procedure.

1. Obtain an initial estimate of β by using ordinary least-squares.
2. Calculate the squared residuals, $r_i^2 = (y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the fitted value for the ith unit, based on the current estimate of β.
3. Estimate γ, by fitting a model in which the r_i^2 are treated as if they were values of a response variate, distributed as $\sigma_i^2 \chi_1^2$. This entails fitting a gamma model with the dispersion parameter value set to 2. The fitted values from this model, $\hat{\sigma}_i^2$, are saved.
4. Re-estimate β by weighted least-squares, with weights of $1/\hat{\sigma}_i^2$.
5. Repeat steps 2-4 until convergence.

This 'see-saw' algorithm is more convenient and efficient than a general optimization procedure.

We now give a simple procedure `D1FIT` which implements this algorithm.

```
PROCEDURE 'D1FIT'
 OPTION 'MAXITER','TOLERANCE'; DEFAULT=15,0.001
 PARAMETER 'Y', 'M_MODEL', 'D_MODEL', 'FITMEAN', 'FITVAR'
   CALCULATE NUnit = NVAL(Y)
   VARIATE MeanWt,Fitm,Fitv; VALUES=!(#NUnit(1)),!(#NUnit(*)),!(#NUnit(*))
   IF UNSET(FITMEAN)  :ASSIGN Fitm; FITMEAN  :ENDIF
   IF UNSET(FITVAR)   :ASSIGN Fitv; FITVAR   :ENDIF
   SCALAR OldLogLik; VALUE=999
   PRINT '  Iteration      LogLik'
   FOR Iteration=1...MAXITER
     MODEL [SAVE=mmodel; DISPERSION=1; WEIGHT=MeanWt] Y; FITTED=FITMEAN
     FIT [PRINT=*; NOMESSAGE=leverage,residual] #M_MODEL
"1 ---->"        "blank"
"2 ---->"         CALCULATE SqrdRes = (Y-FITMEAN)**2
"3 ---->"         MODEL [SAVE=dmodel; DISTRIBUTION=gamma; DISPERSION=2; \
"  ---->"                LINK=log] SqrdRes; FITTED=FITVAR
     RCYCLE [MAXCYCLE=20]
     IF UNSET(D_MODEL)
       FIT [PRINT=*; NOMESSAGE=leverage,residual]
      ELSE
       FIT [PRINT=*; NOMESSAGE=leverage,residual] #D_MODEL
     ENDIF
"4 ---->"        "blank"
      CALCULATE MeanWt=1/FITVAR
      &         LogLik=SUM( (Y-FITMEAN)**2/FITVAR+LOG(FITVAR)+1.8378771)
      PRINT [IPRINT=*; SQUASH=yes] Iteration,LogLik; DECIMALS=0,2
      IF (ABS(OldLogLik-LogLik)<TOLERANCE*ABS(OldLogLik=LogLik))
        RDISPLAY [SAVE=mmodel; PRINT=model,estimates]
        &        [SAVE=dmodel; PRINT=model,estimates]
        EXIT [CONTROL=procedure]
      ENDIF
    ENDFOR
    PRINT [IPRINT=*] '* * *  W A R N I N G   No convergence after', \
          MAXITER,' iterations * * *'; DECIMALS=0; FIELDWIDTH=1
    RDISPLAY [SAVE=mmodel; PRINT=model,estimates]
    &        [SAVE=dmodel; PRINT=model,estimates]
ENDPROCEDURE
```

The response variate is supplied as the `Y` parameter. Model formulae for the mean and dispersion submodels are supplied as the parameters `M_MODEL` and `D_MODEL` respectively. If `D_MODEL` is unset, constant variance is assumed. The parameters `FITMEAN` and `FITDISP` can be used to save the final fitted values for the mean and dispersion submodels. The algorithm terminates when the percentage change in the quantity (*–2loglik*) is sufficiently small (option `TOLERANCE`). For fitting the gamma model at step 3, the maximum number of iterations is set to 20 instead of the default 10, because convergence is sometimes slow, [1]. The markers `"1 ---->"`, `"2 ---->"`, etc. indicate lines of code which will be modified in subsequent sections.

3.2. Methods Based on Other Transformations of the Residuals

Although the use of the squared residuals r_i^2 as a response variable for fitting the dispersion submodel leads to maximum-likelihood estimates when the data are Normally distributed, other transformations of the residuals may be preferable when the data are non-Normal, [4]. Here we consider two possibilities – the absolute residuals and their logarithms.

Use of the absolute residuals is motivated by robustness considerations. It can be much more efficient than use of the squared residuals when the data are from a distribution that is longer tailed than the Normal, and is reasonably efficient even when the data are Normally distributed, [4]. In contrast, use of the logarithm of the absolute residuals is motivated by computational convenience, since it allows the dispersion submodel to be fitted by ordinary least-squares. Statistically there may be problems with units that have small residuals in the mean submodel; when log-transformed, these will become outliers in the dispersion submodel, [4].

We distinguish between the residuals $r_i = y_i - \hat{y}_i$ and the 'errors' $e_i = y_i - \mu_i$. Assuming Normality, we have $e_i^2 \sim \sigma_i^2 \chi_1^2$. In the algorithm of the previous section, the squared residuals r_i^2 are treated as if they had the same statistical properties as the squared errors e_i^2. This suggests that to obtain comparable results with the absolute residuals $|r_i|$ we should treat these as having the same statistical properties as the absolute errors $|e_i|$.

Under Normality, we have

$$E(|e_i|) = \sqrt{2/\pi}\ \sigma_i$$

and

$$VAR(|e_i|) = \sigma_i^2(1-2/\pi)$$

This suggests the following (quasi-likelihood) method of estimating $\frac{1}{2}\gamma$ since $\sigma_i = \sqrt{\exp(z_i^T\gamma)} = \exp(\frac{1}{2}z_i^T\gamma)$: use $\sqrt{2/\pi}|r_i|$ as the response variate and set `DISTRIBUTION=gamma`, `LINK=log`, `DISPERSION=0.5708` ($=\pi/2-1$).

To implement this as a procedure, `D2FIT`, the following changes to `D1FIT` are needed:

```
"2 ---->"       CALCULATE AbsRes = 1.25331 * ABS(Y-FITMEAN)
"3 ---->"       MODEL [DISPERSION=0.5708; DISTRIBUTION=gamma; LINK=log; \
"  ---->"             SAVE=dmodel] AbsRes; FITTED=FITVAR
"4 ---->"       CALCULATE FITVAR = FITVAR * FITVAR
```

A well-known alternative to fitting a gamma model with log link is to take a log transformation of the 'data' (here the absolute residuals) and estimate parameters by ordinary least-squares, [6]. Under Normality,

$$\begin{aligned} E[\log(|e_i|)] &= \log(\sigma_i) - \tfrac{1}{2}[\log(\tfrac{1}{2}) - \psi(\tfrac{1}{2})] \\ &= \log(\sigma_i) - 0.6352 \end{aligned}$$

and

$$VAR[\log(|e_i|)] = \psi'(\tfrac{1}{2})/4 = 1.2337$$

where $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively. (These results follow from, for example, formulae in Exercise 6.13, page 191 of [7].) This suggests the following (least-squares) method of estimating $\frac{1}{2}\psi$: use $\log(|r_i|) + 0.6352$ as the response variate and set `DISTRIBUTION=normal`, `LINK=identity` and `DISPERSION=1.2337`.

To implement this as a procedure, D3FIT, the following changes to D1FIT are needed:

```
"2 ---->"       CALCULATE LogAbsRes = LOG(ABS(Y-FITMEAN)) + 0.6352
"3 ---->"       MODEL [SAVE=dmodel; DISPERSION=1.2337; LINK=identity] \
"  ---->"             LogAbsRes; FITTED=FITVAR
"4 ---->"       CALCULATE FITVAR = EXP(2*FITVAR)
```

3.3. Allowing for Loss of Degrees of Freedom

Section 3.1 showed how maximum-likelihood estimates of the parameters β and γ could be obtained, assuming the data to be Normally distributed. However, maximum-likelihood estimates of the parameters of the dispersion submodel make no allowance for the loss of degrees of freedom associated with the estimation of the parameters of the mean submodel. In the simplest situation of constant variance $\sigma_i^2 = \sigma^2$, the maximum-likelihood estimate of σ^2 is the residual sum of squares divided by n (the total number of observations) instead of by the usual $n-p$ (the number of residual d.f.).

By incorporating the effects of leverage, [4], we can modify the method based on squared residuals to allow for loss of degrees of freedom. For if the data are nearly Normally distributed we have the following approximations:

$$E(r_i^2) \simeq (1-lev_i)\sigma_i^2$$

$$VAR(r_i^2) \simeq 2(1-lev_i)^2\sigma_i^4$$

where lev_i is the leverage of the ith unit. A new procedure, D4FIT, based on these relationships, is got by making the following changes to D1FIT:

```
"1 ---->"       RKEEP LEVERAGE=lev
"2 ---->"       CALCULATE SqrdRes = ((Y-FITMEAN)**2) / (1-lev)
```

Methods based on other transformations of the residuals can be similarly modified. For example, a procedure D5FIT which uses absolute residuals and corrects for the effects of leverage is got by making the following changes to D1FIT:

```
"1 ---->"       RKEEP LEVERAGE=lev
"2 ---->"       CALCULATE AbsRes = Abs(Y-FITMEAN) / SQRT(0.63662*(1-lev))
"3 ---->"       MODEL [DISPERSION=0.5708; DISTRIBUTION=gamma; LINK=log; \
"  ---->"             SAVE=dmodel] AbsRes; FITTED=FITVAR
"4 ---->"       CALCULATE FITVAR = FITVAR * FITVAR
```

3.4. Statistical Problems

This article aims to show how various methods of modelling variance heterogeneity in regression can be implemented in Genstat. However, although the methods are usually computationally straightforward, there are several statistical problems that have not yet been fully resolved. We mention three.

1. Should the see-saw algorithm always be iterated to convergence? It is not obvious that this is always necessary, [2], [4].
2. How should inferences be made about the mean submodel? The mean submodel is estimated by a weighted analysis in which the weights are themselves estimated, but standard errors produced by Genstat strictly assume that the weights are known *a priori*, [2], [12].
3. How should inferences be made about the dispersion submodel? One problem is that assumptions about the third and fourth moments of the data are needed. But even when the data are assumed to be Normally distributed, likelihood-ratio tests, score tests and Wald tests, which are all asymptotically equivalent, can nonetheless give rather different results in small samples, [1].

4. An Example

As an example of multiple linear regression, Chapter 8 of the Genstat 5 Reference Manual uses some data from [5] in which the response variate is the monthly consumption of water (thousands of gallons) at a plant for which four possible explanatory variates are available:

Temp – average monthly temperature (degrees F)
Product – amount of production (billion pounds)
Opdays – number of plant operating days in the month
Employ – number of people employed

The data are for 17 consecutive months, but this ordering is ignored in the following analysis which is intended to be illustrative, not definitive.

The parameter estimates from fitting the model `Temp+Product+Opdays+Employ` are given in Table 1 and the residual mean square is 0.0620 (12 d.f.).

Parameter	Estimate	s.e.	t
Constant	6.36	1.31	4.84
Temp	0.01387	0.00516	2.69
Product	0.2117	0.0455	4.65
Opdays	–0.1267	0.0480	–2.64
Employ	–0.02182	0.00728	–3.00

Table 1

Residual plots, [3], suggest that the variability may be greater in warmer months. The estimates for the dispersion submodel from using the procedures `D1FIT-D5FIT` with `D_MODEL=!F(Temp)` and `M_MODEL=!F(Temp+Product+Opdays+Employ)` are given in Table 2.

	Procedure	Type of Residual	Adjustment for Leverage	Coefficient of *Temp* in dispersion submodel		
				Estimate	s.e.	t
	`D1FIT`	Squared	No	0.157	0.0262	6.02
†	`D2FIT`	Absolute	No	0.201	0.0280	7.18
†	`D3FIT`	Log-Absolute	No	0.282	0.0412	6.85
	`D4FIT`	Squared	Yes	0.118	0.0262	4.27
†	`D5FIT`	Absolute	Yes	0.121	0.0280	4.32

Table 2

For the procedures marked with †, the estimates and standard errors given by Genstat have been multiplied by 2 so that all estimates in the table are of the same parameter (see Section 3). The estimates from procedures `D1FIT`, `D2FIT` and `D3FIT` are dissimilar. Limited experience with other data sets suggests that estimates from `D1FIT` and `D2FIT` are usually more similar than in this example but that estimates from `D3FIT` may differ substantially. However, as in this example, t-values from the three procedures are often more consistent.

When an allowance is made for leverage there is, in this example, little difference between using squared residuals (`D4FIT`) and using absolute residuals (`D5FIT`). The t-values for all the procedures suggest that the apparent increase in variability with increasing *Temp* is statistically significant (even if the true significance level is much less than is suggested by the t-values). The corresponding likelihood ratio statistic, assuming Normality, is 8.19 (1 d.f.), again highly significant.

The effects of this non-constant variance on the parameter estimates for the mean submodel are reasonably consistent for the different procedures. For example, Table 3 shows the estimates from procedures `D1FIT` and `D4FIT`.

Parameter	Estimate			t-value		
	Least-squares	D1FIT	D4FIT	Least-squares	D1FIT	D4FIT
Constant	6.36	5.10	5.44			
Temp	0.0139	0.0093	0.0106	2.69	2.24	2.41
Product	0.212	0.187	0.194	4.65	6.75	6.11
Opdays	–0.127	–0.0722	–0.0883	–2.64	–2.27	–2.38
Employ	–0.0218	–0.0181	–0.0189	–3.00	–5.73	–4.83

Table 3

The regression coefficients are all less in absolute value than the ordinary least-squares estimates. Standard errors are also reduced, as is to be expected because of the weighting, but the reductions are not consistent for all four variables. Instead, the effects of *Product* and *Employ* are now estimated with relatively more precision (i.e. they have greater t-values) than in ordinary least-squares, whilst the relative precision with which the effects of *Temp* and *Opdays* are estimated is slightly reduced.

Finally, prediction is one aspect of regression that is particularly sensitive to assumptions about variance structure. For illustration, Table 4 gives predicted water usage at five temperatures, with the other variates (*Product*, *Opdays*, *Employ*) fixed at their mean values within the original data set. The predictions are based on the parameter estimates obtained by ordinary least-squares and by procedure `D4FIT`, and were produced by the command:

```
PREDICT Temp; LEVELS =!(40,50...80)
```

Standard errors produced by the `PREDICT` command are appropriate when the predictions are regarded as predictions of mean response at the given temperature. The standard errors given in the Table are appropriate when the predictions are regarded as predictions of individual observations. They are got by squaring the standard errors from `PREDICT`, adding the variance of an individual observation (which is constant for least-squares but depends on *Temp* for `D4FIT`) and taking the square root.

Temp	Prediction		Standard error	
	Least-squares	D4FIT	Least-squares	D4FIT
40	2.96	3.19	0.287	0.061
50	3.10	3.29	0.267	0.085
60	3.24	3.40	0.257	0.154
70	3.38	3.50	0.258	0.266
80	3.51	3.61	0.268	0.453

Table 4

The predictions produced by the two methods, and more particularly the standard errors of the predictions, are very different.

5. Extensions

This article has discussed parameter estimation for models in which the variance is functionally independent of the mean. The Genstat procedures assume the submodel for the mean to be an ordinary linear model, but they are easily modified to cope with other link functions. The methods are equally applicable to general nonlinear models for the mean, though this would require more substantial changes to the procedures.

Similar methods have been suggested for generalized linear models, particularly in connection with modelling overdispersion in discrete data. Smyth, [10], discusses maximum-likelihood methods for models with non-constant dispersion parameter when the data are assumed to be gamma or inverse-Normally distributed, thus generalizing the results of Section 3.1.

For GLMs, residuals can be defined in several ways. For example, extended quasi-likelihood, [8], [9], is based on the squared deviance residuals, with no adjustment for leverage, whilst Williams, [13], suggests using the squared Pearson residuals and allowing for the effects of leverage.

The advantages and disadvantages of the various methods are not yet completely understood. But computationally they can all be implemented using a see-saw algorithm which alternates between fitting the mean submodel with weights based on current estimates from the dispersion submodel, and fitting the dispersion submodel with the response variate based on the residuals from the mean submodel. This in turn implies that the methods are easily programmed using the existing regression facilities in Genstat.

6. References

[1] Aitkin, M.
Modelling Variance Heterogeneity in Normal Regression using GLIM.
Applied Statistics, 36, pp. 332-339, 1985.

[2] Carroll, R.J., Wu, C.F.J. and Ruppert, D.
The Effect of Estimating Weights in Weighted Least-squares.
Journal of the American Statistical Association, 83, pp. 1045-1054, 1988.

[3] Cook, R.D. and Weisberg, S.
Diagnostics for Heteroscedasticity in Regression.
Biometrika, 70, pp. 1-10, 1983.

[4] Davidian, M. and Carroll, R.J.
Variance Function Estimation.
Journal of the American Statistical Association, **82**, pp. 1079-1091, 1987.

[5] Draper, N.R. and Smith, H.
Applied Regression Analysis.
Wiley, New York, (2nd edition) 1981.

[6] Firth, D.
Multiplicative Errors: Log-normal or Gamma?
Journal of the Royal Statistical Society, Series B, **50**, pp. 266-268, 1988.

[7] Kendall, M.G. and Stuart, A.
The Advanced Theory of Statistics, Volume 1: Distribution Theory.
Charles Griffin and Company, London and High Wycombe, (4th edition) 1977.

[8] Nelder, J.A.
Overdispersion and Extended Quasi-likelihood.
Proceedings of the XIVth International Biometric Conference, pp. 289-300, 1988.

[9] Nelder, J.A. and Pregibon, D.
An Extended Quasi-likelihood Function.
Biometrika, **74**, pp. 221-232, 1987.

[10] Smyth, G.K.
Generalized Linear Models with Varying Dispersion.
Journal of the Royal Statistical Society, Series B, **51**, pp. 47-60, 1989.

[11] Stirling, W.D.
Heteroscedastic Models and an Application to Block Designs.
Applied Statistics, **34**, pp. 33-41, 1985.

[12] van Houwelingen, J.C.
Use and Abuse of Variance Models in Regression.
Biometrics, **44**, pp. 1073-1081, 1988.

[13] Williams, D.A.
Extra-binomial Variation in Toxicology.
Proceedings of the XIVth International Biometric Conference, pp. 301-313, 1988.

Use of Genstat at the International Maize and Wheat Improvement Centre

C Gonzalez
CIMMYT
Apartado Postal 6-641
06600 Mexico, D.F.
Mexico

The International Maize and Wheat Improvement Centre (CIMMYT) is one of 13 non-profit international research centres supported by the Consultative Group on International Agricultural Research. CIMMYT's mandate is to support the maize and wheat research work of national programs throughout the world, with emphasis on the production problems of developing countries.

At CIMMYT headquarters in Mexico there are over 100 scientists, research students and assistants who conduct about 300 wheat and maize variety and agronomic trials in Mexico, and distribute wheat yield nurseries and maize variety trials to about 400 locations around the world. There are also 45 CIMMYT scientists posted in 18 regional offices in Africa, Asia and Latin America working in close collaboration with the national agricultural research services.

The statistical needs of CIMMYT's scientific community are serviced by a Statistics Unit which has grown in strength from nought to two professionals in the last six years. To cope with the workload, the Unit requires that scientists be responsible for a large part of their own computing.

There are thus, at CIMMYT, four groups of people whose needs and skills with regard to computing and statistics vary greatly:

(a) non-statisticians who in addition to generating designs and handling the maize and wheat international trials are responsible for analysing the resulting data;

(b) non-statisticians responsible for the statistical analysis of their own experiments;

(c) non-statisticians unskilled in statistics and/or computing who hand their data over to the Statistics Unit for analysis; and

(d) the Unit's professional statisticians who, besides analysing other people's data, need to perform more sophisticated analyses than the first two groups.

Our aim was to provide all four groups with a single statistical package that could be used by statisticians as a computing language for performing sophisticated analyses, as well as by non-statistically-minded scientists for simple data validation, tabulation, graphing and routine statistical analyses.

Genstat, which in our opinion fulfilled the above conditions, became available at CIMMYT at the beginning of 1984. Since then it has been extensively used by the Unit's statisticians for analysing individual trials and yield stability across environments using the linear regression approach and a geometric method based on a principal co-ordinates analysis of a matrix of similarities between genotypes.

However, many of CIMMYT's non-statisticians, who are occasional users unskilled in statistics and computing, found Genstat 4 too difficult to learn. So they either handed their data over to the Statistics Unit or used SAS instead.

Since the release of Genstat 5 seemed to be imminent, we at CIMMYT's Statistics Unit did not push too hard for people to learn to use Genstat 4. It was not until the beginning of 1989 that formal training on Genstat started at CIMMYT. As a result, it is becoming increasingly popular among non-statisticians.

A set of two Genstat courses, an introductory course and an advanced one, have been given at CIMMYT during 1989, to nine staff members from the Maize and Wheat programs and support services. Attendees have been selected for their involvement in data analysis within their respective sections. In both courses short lectures and practical classes are given alternately.

During the three-day introductory course, the students are taught to handle data, produce graphs and histograms and tabulate data. Over the two days of the more advanced course, the students are taught to analyse designed experiments using ANOVA and perform regression analysis on continuous and grouped data. Before teaching the students how to run an ANOVA on Genstat, a few hours are needed to make them understand the distinction between the two components of an experimental design – the structure of the experimental units and the structure of the treatments.

I think formal training is important, but preferably if the prospective student has a positive attitude towards Genstat. The statistician can help bring this about by holding consultation sessions and highlighting some of the features where Genstat outperforms other packages, e.g. multiple error terms in the ANOVA, operations with tables, generalized linear models and the possibility of saving output structures for future use.

As a result of the simplified and more consistent syntax of Genstat 5 – which has made it easier to learn and use – of its truly interactive mode and its improved documentation, novices tend to react to it much more positively than to Genstat 4.

The need to generate designs, construct experimental plans, and analyse individual variety trials arranged in alpha lattice designs, as well as unbalanced variety by site and variety by season tables of mean effects, has led to an increased use of special software such as DSINGX, ALPHAGEN, ALPHANAL and REML, distributed by the Scottish Agricultural Statistics Service.

As a result of a growing concern at CIMMYT for extracting more information from series of variety trials, REML has become by far the most intensively used of the statistical programs mentioned above. However, because the current version of REML is inflexible in its input and much more so in its output, it is inadequate for producing reports. Adding REML facilities to Genstat will overcome this inconvenience, and will make it possible for the group of people involved in analysing series of variety trials to become part of CIMMYT's Genstat users' community.

CIMMYT's outreach officers around the world use PCs for analysing experiments they conduct with agricultural research program staff of the regions where they are posted. Most of them have been using MSTAT or SYSTAT for their analyses. They do not contemplate licensing SAS because of the price, nor Genstat because of the difficulty of getting any kind of training or consultation support from headquarters. They too, as well as their partners in the national agricultural research programs, are likely to benefit from a conversational interface.

Data from CIMMYT maize and wheat trials have not yet been organised in databases. Therefore, for most types of analyses, it takes much longer to gather and organise the data than to write the code (Genstat or other) and run the job. For example, because the mean effects of varieties from individual trials are not stored, in order to analyse a site by variety table of mean effects, it is necessary to re-analyse the individual trials and save the mean effects to construct the table that finally will be analysed using REML.

We hope in the not too distant future to have data organised in databases, and then to develop an interface between the databases and Genstat, so that the user need not bother with the mechanics of transferring data from the database to Genstat.

Summing up, we foresee an important increase in the use of Genstat by CIMMYT's non-statisticians, both on the mainframes and on PCs as a consequence of:

- the simplified and more consistent syntax of Genstat 5, which has made it easier to learn and use;
- its truly interactive mode;
- its improved documentation;
- the future addition of REML and a conversational interface; and
- the availability of procedures to 'customise' Genstat for a wide range of applications, and of an interface between databases and Genstat.

Experience with Genstat in Teaching an Applied Statistics Course

C Donnelly
Australian National University
GPO Box 4
Canberrra City
ACT 2601
Australia

1. Background

'Statistics for Research Workers' is a one-semester course in the application of statistics to data analysis that the Statistics Department at the Australian National University offers as part of its applied statistics program. In 1989, the course attracted an enrolment of about 40, made up of research workers, both staff and postgraduate students (auditing) and third year undergraduates, formally enrolled. The majority of students were from science disciplines.

With such diversity in student types, there was considerable variation in student background. Computing experience ranged from one student with a computer science major to several who had never used a computer of any kind. A range of style of computing was also represented, with some students familiar with the Macintosh interface and others, command-driven systems, either mainframe or PCs. Correspondingly, a wide range of statistical experience was represented, from no formal tuition up to five mathematical statistics courses. Experience with computing did not guarantee any previous experience with statistical packages.

2. Methods of Teaching

The method we adopted of teaching by example aimed at providing students with sufficient knowledge of Genstat 5 to solve the problems presented to them, rather than a comprehensive knowledge of the full power and range of options provided by the package.

We rejected using the 'reference manual' style approach as being too daunting to even the most dedicated student. Having to wade through detailed descriptions of a large range of options and parameters associated with each directive is not only time-consuming and confusing, but unnecesary to the novice with a simple, straightforward task at hand. This does not mean that such options and parameters are unnecessary; they provide the language with the flexibility of use that gives it its power.

From the very first lecture, students were provided with examples of Genstat 5 code to perform the tasks under discussion. Tutorial examples were set, starting with very simple commands for the basic tasks of entering and manipulating data. Plotting techniques, factor specification and the powerful TABULATE directive were also introduced using examples. We illustrated how to extend or modify the effect of Genstat 5 directives by using options and parameters. The use of procedures was also found to be essential in our teaching environment, enabling students to make use of, for example, diagnostic plotting techniques without having to write the Genstat 5 code themselves. One of the major advantages this has is that it allows the students to feel more in control of the data and analysis being performed. We also introduced the concept of 'extended interactive' mode, where commands are read from an existing file and executed interactively. This was found to be a powerful teaching tool, saving time and providing the students with material to work through later at their own pace and to use as examples for their own programs.

By the end of 12 weeks, students were able to write their own Genstat 5 code to manipulate and display data, and to perform regression analyses or analyses of variance with a range of displays for diagnostic checking. A practical examination was set as part

of the course assessment. Students were given two hours to write and execute the appropriate code to solve two problems:

– a multiple regression, with outliers;

– an analysis of variance, requiring transformation of the response variable.

Only three of the 26 students who attempted the examination failed. This failure rate was no greater than that for a previous course using Minitab and was much less than the failure rate in the written statistics exam. It would seem that difficulties students experienced in understanding the statistical concepts presented to them during the course were not confounded by an inability to learn and use Genstat 5.

3. Conclusion

Despite expressions of negative feelings from colleagues, we believe Genstat 5 offers a suitable computing environment for teaching applied statistics courses. Its obvious strengths from our viewpoint are its flexibility, the wide range of procedures and an attractive modelling philosophy for teaching statistics.

Experiences with Genstat 5 on Personal Computers

V van den Berg
Institute of Agricultural Engineering
Wageningen
The Netherlands

1. Introduction

Prior to the introduction of the PC version of Genstat 5 Release 1.2, by the Ministry of Agriculture and Fisheries in the Netherlands, I have had the opportunity to work with this version. This memorandum is a brief report on my experiences; moreover it contains a few recommendations from which future users may benefit.

2. Restrictions

The memory capacity required for Genstat on a PC is 581 Kb RAM (see also Section 5). The general character workspace can contain 32,768 characters. The general numerical workspace can contain a total of 39,000 reals; of these over 16,000 are occupied by 'system'. This means that the remaining capacity for the user is nearly 23,000 reals. This might be rather low, and the use of the DELETE directive appears to be necessary quite often. In Section 4 an alternative solution is proposed to solve this problem.

3. Comparison of CPU Time

The CPU times for three different Genstat programs have been compared using a MicroVAX 3500, a MicroVAX II (both with the VMS 4.7a Operating System), an Olivetti M280 (AT) and an Olivetti M240 (XT) (both with MS-DOS version 3.20 and mathematical co-processor).

	MicroVAX		Olivetti PC	
Genstat program	3500	II	M280	M240
Linear regression analysis, 20 observations	3	8	115	161
Program with many calculations (including IFs and FORs), 2 datasets, 25 observations each	13	40	1050	1410
Nonlinear regression analysis, (including use of own procedure), 50 observations	25	80	1420	1875

Table 1

Comparison of CPU times (in seconds) on different computers

4. Adaption of Language Definition File

My attention was drawn to the possibility of extending the capacity available for data structures. In the Language Definition File all Genstat directives are defined. It is conceivable that not all these directives need to be available (all the time). If a number of directive definitions in the Language Definition File are omitted, the system occupies less of the total numerical workspace, and so more capacity becomes available to be used for data structures. In the shortened version of the Language Definition File as dealt with below, a number of directives referring to time-series analysis and cluster analysis have been omitted. The result is shown in Table 2. The procedure is further described in Appendix 1.

Version	System	Data structures	Total
Original	16,000	23,000	39,000
Shortened	10,000	29,000	39,000

Table 2

Capacity available for Language Definition File

5. Use of Memory

In many cases drivers will be installed in a PC and programs are resident in the memory (such as CED, Deskmate, Pctools, Gmouse, etc.). As a result the memory capacity available will be (far) below the required capacity of 581 Kb. The presence of an extended/expanded memory does not solve this problem because in Genstat itself, no povisions have been incorporated for this; and MS-DOS (certainly up to Version 3.30) cannot support more than 640 Kb of memory. In the next section it is indicated how use of extended memory can be made for another purpose.

To have sufficient memory capacity, and not to do without the convenience of resident programs, a series of batch procedures have been developed, which take care that these programs are switched off temporarily, i.e. only when Genstat is running.

These procedures are:

`MSDOSMEM.BAT` with which the usual programs are made resident in memory;

`GENSTMEM.BAT` with which the resident programs can be removed from memory and with which a simple CED version can be installed in memory (with only the cursor keys and the `G5` command being defined);

`G_START.BAT` with which `GENSTMEM.DAT` is started (this batch file is to be run before starting Genstat);

`G_END.BAT` with which `MSDOSMEM.BAT` is started (this batch file is to be run after termination of one or several Genstat programs).

`MSDOSMEM` and `GENSTMEM` use the `MARK.COM` file, marking the spot from where programs are to be removed from the memory. `G_START` and `G_END` use the file `RELEASE.COM`, removing the programs from the memory. In addition, it is useful to have `MAPMEM.COM`, which indicates how much capacity is occupied by resident programs.

The use of Genstat could then be as follows:

```
G_START
 G5 j0b1.gst
 G5 j0b2.gst
 G5 j0b3.gst
 etc.
G_END
```

Appendix 2 shows the contents of the above auxiliary files. Additionally, an example is given of the matching `AUTOEXEC.BAT` file.

6. Use of Extended Memory

If more than 3 Mb of extended memory is available on the PC, the speed of Genstat can be raised considerably by using virtual disk capacity. For that reason the extended memory is declared to be a 'disk' by means of the `VDISK` DOS command. Instead of using Genstat from the hard disk, this is done from the memory. Considerable time can be gained, especially for reading and writing.

6.1. Example of Virtual Disk Use

Assuming the presence of a total memory capacity of 4 Mb, or rather 640 Kb of conventional and 3456 Kb of extended memory. The extended memory can then be defined as a virtual disk by inserting the following line in the CONFIG.SYS file (assumed to be present in the C:\BIN directory):

```
DEVICE = C:\BIN\VDISK.SYS 3456 /E
```

The virtual disk then automatically gets the next drive letter that is not yet in use. So, if there is a floppy-disk drive called A: (and possibly also one called B:) and a hard disk designated C:, the virtual disk drive letter will be D:.

Two more actions are required for working with Genstat:

- Genstat files have to be copied to D: and
- in the AUTOEXEC.BAT file the line

```
SUBST G: C:\GENSTAT
```

has to be replaced by:

```
SUBST G: D:
```

To indicate the gain in CPU time by using a virtual disk, the same three programs as presented in Table 1 have been used with an Olivetti M280 *with* and *without* using the virtual disk (Table 3).

For these programs, the use of an extended memory as a virtual disk results in a CPU time gain of approximately 40%.

7. Use of PC-Cache

Another suitable MS-DOS feature to raise the speed of Genstat on the PC is PC-Cache. The disk cache system stores frequently accessed backing store data in a main memory buffer (i.e. cache). This means that when this data is read in by a program, it is read from the main memory instead of from backing store. This can improve the access time very much. You can included the PC-CACHE command in your AUTOEXEC batch file, e.g.

```
PC-CACHE /IA /IB /SIZEXT= 512
```

For the meaning of the parameters, see your MS-DOS manual.

Table 3 compares the CPU time for the three Genstat programs. The use of PC-Cache gives another slight reduction of CPU time.

Genstat program	Hard disk	Virtual disk	PC-Cache
Linear regression analysis	115	79	74
Program with many calculations	1050	650	520
Nonlinear regression analysis	1420	1020	910

Table 3

Comparison of CPU time (in seconds), running Genstat on a hard disk, on a virtual disk and with PC-Cache

8. Conclusions

The practice of working on a PC with Genstat 5 Release 1.2, can be summarised as follows:

- the PC version is practically identical with the (Micro)VAX version of Genstat 5 (with only the high-quality graphics and Fortran extension facilities lacking), so that

one can almost directly start using it; no new syntax has to be learned and one can use the same manuals;

- this also implies that with a PC one can dispose of a program of the same high statistical quality as Genstat 5;
- only small jobs can be managed within an acceptable CPU time; for bigger jobs a MicroVAX will remain the proper machine. If, however, use can be made of adequate virtual disk capacity, a substantial reduction in CPU time can be accomplished; another reduction can be gained by using the PC-Cache feature of MS-DOS;
- it has appeared to give quite an amount of trouble to combine adequate memory capacity and user-friendly PC properties; reports from Rothamsted lead me to expect that (a part of) these problems will be solved in the next release.

9. Reference

[1] Harding, S.A. and Trinder, K.I.
Features of the Genstat 5 Language: 1.
Genstat Newsletter, 22, pp. 51-55, 1988.

Appendix 1: Description of the Language Definition File Adaptation

- Go to the subdirectory of which Genstat is installed; default `C:\GENSTAT`
- Copy the file `G5DEFS.TXT` to (e.g.) `ORIGDEFS.TXT`
- Rename the file `G5BOOT.BIN` to (e.g.) `ORIGBOOT.BIN`
- Edit `G5DEFS.TXT` and remove (or put in quotation marks " ") those directives one can do without
- Run Genstat once without a program (i.e. only give command `G5`); a new version of `G5BOOT.BIN` is now made (an Olivetti M280 machine needs approximately 18 minutes for this)
- The shortened version can now be used

Remark: A complication might arise in the above procedure in that a message 'unknown directive' appears when a Genstat program is used, while this directive is (still) included in the `G5DEFS.TXT` file. The reason is that some directives assume that others have been defined. This can be solved by putting back in the file those directives used by the directives left. It is indicated in Genstat Newsletter 22, page 55, [1], which directives are involved.

Appendix 2: List of Auxiliary Files Used with Genstat on a PC

`MSDOSMEM.BAT:`

```
echo off
mark MSDOS
ced -fc:\bin\ced.cfg          } example of resident programs switched off during use
c:\deskmate\deskmate /m       } of Genstat
cls
```

`GENSTMEM.BAT:`

```
echo off
mark GENST
ced -fc:\bin\ced.gen          - only CED.GEN remains resident (in it, G5 is defined)
cls
```

CED.GEN:

```
syn g5 c:\genstat\genstat
```
– definition of symbol G5

G_START.BAT:

```
echo off
release MSDOS                 – resident programs out of memory
genstmem                      – G5 definition is loaded
cls
```

G_END.BAT:

```
echo off
release GENST                 – CED.GEN out of memory
msdosmem                      – resident programs back in memory
cls
```

AUTOEXEC.BAT (example):

```
echo off
path c:\;c:\bin;c:\spss;c:\pico;c:\pap\pae;c:\usr;c:\genstat
prompt $p$g
subst g: c:\genstat
etc.
msdosmem
```

Genstat 5 Release 1 for Personal Computers

P G N Digby
Statistics Department
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
Harpenden
Herts AL5 2JQ

1. Introduction

This article is intended as a companion to that by Valentijn van den Berg, 'Experiences with Genstat 5 on Personal Computers'. The version of Genstat 5 described in that article is Release 1.2; at the time of writing this article the implementation of Release 1.3 for PCs has been completed, and it will be available from NAG early in 1990.

2. Release 1.2

Valentijn van den Berg has provided some useful suggestions as to how difficulties with various problems can be overcome, or at least alleviated. I can confirm that the DELETE directive needs to be used fairly frequently, although it is not necessary immediately after a CALCULATE statement, or any multivariate statement, since these automatically do the internal equivalent of the DELETE directive when they have finished. The scheme of reducing the Language Definition File is certainly useful, as evidenced above; however, for Release 1.3 it will not be necessary (see below). Valentijn van den Berg's article suggests a useful method of making sufficient memory available for Genstat 5, although I am not sure whether the programs MARK, RELEASE, and MAPMEM are available in the Public Domain. Unfortunately, we do not have a PC with 4 Mb of memory so I cannot confirm the comments about the use of a 3456 Kb virtual disk.

3. Release 1.3

Our intention with Release 1.3 has been to improve Genstat 5 for PCs, although certain improvements will have to wait until Release 2. I am grateful to many of my colleagues, in particular Roger Payne, Rodger White and Steve Haywood, for their suggestions and comments, and for helping with the implementation.

We saw six aspects of Release 1.2 that needed to be considered:

1. introduction of separate windows for input and output;
2. introduction of facilities for high-resolution graphics;
3. speed of execution;
4. amount of memory needed to run Genstat 5;
5. amount of data space available to the user; and
6. use of extended memory.

In addition we were aware of some users' difficulties in loading the program onto a PC. This is because the program itself, for Release 1.2, consists of a single file which is supplied as an MS-DOS BACKUP file formed using Version 3.2 of MS-DOS. Unfortunately some users do not have Version 3.2 (or 3.3), and some PC manufacturers do not implement the RESTORE command correctly – this is where the problem lies.

3.1. Windowing

We thought that some sort of windowing could be useful, in line with other programs for PCs, although at this stage not to the extent that some programs provide. Whilst the Genstat 5 Committee has several ideas on the use of windowing, e.g. for HELP

information, we have chosen to implement the basic idea of separate windows for input and output for Release 1.3 on PCs.

The user can choose whether windowing is to be used, and must specify the size of the windows, when Genstat 5 is started; by default windowing is not used. The foreground and background colours can also be changed.

3.2. Graphical Facilities

As for other implementations of Genstat 5, the high-resolution graphical directives have been implemented using the Graphical Supplement from NAG: I am grateful to Lesley Carpenter, of NAG, and her colleagues for making this available to me as Fortran source code. I have had to exclude the possibility of producing graphical output on printers, but otherwise the graphical devices available in Release 1.3 are the same as those available in the PC version of the NAG Graphics Library.

If windowing is being used, the graphical display can be directed to the output window and retained there, in which case any other output is directed through the input window. Thus the construction of complicated displays, e.g. involving several DGRAPH statements, is quite possible.

3.3. Execution Speed

The executable program for Release 1.3 has been split into 161 separate files. This has improved the speed so that Release 1.3 takes about 75% of the execution time that Release 1.2 takes. It also removes the difficulty of using MS-DOS BACKUP files, as none of the separate files is too large to fit onto a single 5.25" double-sided double-density (360 Kb) disk.

However, an additional aspect is the way that Genstat 5 is loaded onto the hard disk. If the disk is very fragmented, so that the Genstat 5 files are scattered, and possibly split, over large areas of the disk then it will be very much slower. This is impossible to quantify exactly, since it depends on so many factors; however, Table 1 shows the times taken to run one set of examples on a IBM PS/2 Model 60.

Version	Minutes
Release 1.2 (contiguous)	28
Release 1.3 (contiguous)	21
Release 1.3 (fragmented)	33

Table 1

Therefore it makes sense to ensure that Genstat 5 is implemented as contiguously as possible. From time to time I BACKUP (most of) the files on a disk drive, delete the files, and RESTORE them: this ensures that the files are contiguous on the disk. I would recommend that users of Genstat 5 on PCs do this, either before loading Genstat 5, or afterwards. Alternatively, many users will have a utility which unfragments files and directories.

3.4. Execution Size and Data Space

The difficulties over the space needed to run Genstat 5 are caused by the need to compromise between three things. Genstat 5 on PCs is overlaid, so that only part of the program is resident in memory at any time and other parts must be fetched from the disk as required (this is why the disk-activity light flickers almost continuously whilst using Genstat 5). We could tighten up the overlay, so that less of the program was resident in memory; however, this would increase the time taken enormously, since only a few

much-used subroutines and COMMON blocks remain resident. We have chosen to leave the overlay of subroutines (provided by Howard Simpson) unchanged between Release 1.2 and Release 1.3, apart from the introduction of the high-resolution graphical directives.

The second consideration is the overall memory needed to run the program. Comments on this concern a fairly minimal set of Terminate-and-Stay-Resident (TSR) programs, and the desire to use Genstat 5 on networked PCs.

The third consideration is the amount of data space available to the user. In Genstat 5 Release 1 the binary version of all of the directive definitions remains in the user's data space. For Release 2 this has been changed so that the directive definitions are cached. We have borrowed this idea for Release 1.3 on PCs, so that nearly all of the data space is available to the user; Simon Harding provided the necessary pre-release code. This means that there is nothing to gain from reducing the set of directives available.

Release 1.3 will be available in several forms, with different amounts of space for real values. The smallest version that we have implemented will run in 506 Kb (1 Kb = 1024 bytes), so should work within a networked environment; this version has space for 24,000 real values. The standard version has space for 32,768 real values, and runs in 539 Kb. We have also constructed a version with space for 40,000 real values: this needs 566 Kb and will run on your PC if you already use Release 1.2.

We have decided that the use of extended memory has so many complications, and implications, that it should await Release 2 of Genstat 5. However, it has not escaped us: various possibilities will be under assessment by the time that you read this.

4. Genstat 5 Release 1.3 for 80386-based PCs

The version of Genstat 5 for PCs with an 80386 processor-chip is extremely fast; no doubt some readers will have seen it by the time that this article reaches them. That version has been produced using the combination of an excellent compiler and linker, that are only available for PCs with an 80386 processor-chip. The 80386-version of Release 1.3 takes about one minute to run the same set of examples as used for the values given in Table 1, on a Compaq Portable 386 (80386 + 80387 co-processor, both running at 20 MHz; 1 Mb memory; 1 Mb paging file; SUSPEND command disabled).

Use of Genstat and Other Software in Graduate Students' Problems

P M E Altham
Statistical Laboratory
University of Cambridge
Cambridge
United Kingdom *CB2 1SB*

1. Introduction

The Cambridge University graduate course leading to the Diploma in Mathematical Statistics contains a substantial Applied Project. Each student is assigned a problem, typically arising from another department of the University or from local industry or research units, and supervised by an External Supervisor, usually, but not always, a practising statistician.

The students are required to submit written accounts of their projects before the examinations in June. A list of titles for 1988-89, together with the External Supervisor's name and affiliation, is given in Table 1.

UNIVERSITY OF CAMBRIDGE DIPLOMA IN MATHEMATICAL STATISTICS 1988-89 Applied Projects	
Avery, C.N.	Neural Network Simulations Using the Back Propagation Algorithm. (*Professor P. Whittle, Statistical Laboratory*)
Deans, W.S.	Classification of Back Pain. (*Dr R. Hanka, Dept. of Community Medicine*)
Ellis, J.E.	Pottery and Politics. (*Dr T. Whitelaw, Dept. of Archaeology*)
Gray, A.D.	Adult and Youth Unemployment; A Structural Analyis for the United Kingdom. (*Mr K. Lee, Dept. of Applied Economics*)
Hooper, R.L.	Two Models for the Distribution of Whales in the Antarctic and the Associated Abundance Estimates. (*Mr L. Hiby, British Antarctic Survey*)
Johnson, A.S.F.	Analysis of the Dolleman Ice Core Data (British Antarctic Survey). (*Ms J. Bryan-Jones, British Antarctic Survey and An Teallach Ltd.*)
McNeil, A.J.	Infant Mortality: A Statistical Analysis of 10 English Parishes, 1581-1855. (*Dr J. Oeppen, History of Population and Social Structure*)
Nason, G.P.	Investigation of Ice-Floe Distribution and Orientation. (*Mr A.M. Cowan, Polar Oceans Associates*)
Rooprai, A.	Comparing Techniques to Analyse Interaction Effects in a 2×2 Completely Randomised Design. (*Mr D. Brown, AFRC Institute of Animal Physiology and Genetics Research*)
Satchell, A.S.G.	An Analysis of Human Childhood Growth (Birth to Four Years of Age). (*Dr T.J. Cole, MRC Dunn Nutrition Unit*)
Shippey, T.H.	Analysis of Data from Quantitative Autoradiography. (*Dr A. Davenport, Clinical Pharmacology Unit*)

Table 1

Nine of these students were supported for the course by SERC Advanced Course Studentships, one by his employers, Government Communications Headquarters, and one by a Churchill Fellowship.

Clearly, doing such an Applied Project will be educational to a student for lots of reasons. Readers of the Genstat Newsletter may be interested to see the uses to which Genstat, and other software, have been put by a typical group of 11 students. They have access to Genstat, GLIM, SPSS-X and a variety of other software languages. Summaries of these projects, as written by their student authors, follow.

2. The Projects

2.1. Neural Network Simulations Using the Back Propagation Algorithm

This project concerns pattern recognition in neural networks. From a given string of input information, we want to find a network that will identify that string as possessing or not possessing a given pattern by the means of a series of successive numerical calculations. Current research work in this area centres on the development of deterministic learning rules which can be used successfully to create and calibrate networks to identify particular patterns *without* any human intervention. The Back Propagation Algorithm is the cause of some recent disagreements. Its proponents claim that it is nearly always successful and can be used to solve complex problems in huge neural networks. However, Minsky and Papert, two of the original theoreticians in the related field of simple perceptrons, dispute such claims as extravagant, arguing that extensive simulations will disprove the efficacy of the algorithm. The purpose of this project is to try to build an understanding of the algorithm's potential through simulations regarding (1) its reliability and (2) its ability to produce original solutions to problems which do not offer an immediate solution. From the fairly simple problems we tackle, the algorithm performs well on both grounds, but not to the standard claimed by its supporters. Most of the algorithm's failures are due to an indirectness in its evaluation of the error inherent in a given network. Improving the algorithm's definition of network error should be the first priority for any researchers who intend to argue for back propagation as an ideal network training algorithm. The computation was carried out mainly by specially written Pascal programs, and partly by using a spread-sheet program.

2.2. Classification of Back Pain

The diagnosis of back pain is currently one of the problem areas in medicine. Back pain is extremely common, and there are many known pathologies which can produce symptons of pain. However, in most cases it is very difficult to find a causal relationship between pain as experienced by the patient and a pathology or set of pathologies as understood by a doctor.

One reason for this is the complex nature of the spinal column, and the variety of stresses applied to it during the life of the patient. These stresses are so severe that they cause physical defects in the various parts of the spinal structure. Such defects are common throughout the population, even among patients who are not aware of any pain, so that it is not possible to have a simple causality such as: pathology *A* leads to pain felt at site *X*.

The difficulty of diagnosis has led to the evolution of various schools of practice, generally associated with different areas of specialisation, diagnosing on the basis of different clinical observations.

The matter is made more difficult by the subjective nature of many of the syndromes proposed for use in classifying back pain. These have been found clinically useful by the doctors who proposed them, but little research has been done to transfer them to a sounder, less subjective basis. This is not to say that they are diagnostically unsound,

but that they are not reliable concepts for the transmission of information throughout the medical community. Two different practitioners wil have different ideas of the interpretation of a particular syndrome, and therefore will be unable to communicate in a common language.

The first step in the eventual diagnosis of back pain is therefore to agree on a description of the symptoms. It is necessary that a group of doctors agree on this before studies can be made over large numbers of patients from different clinics, seen by different members of the group.

The Birmingham questionnaire, formulated by Mr Jeremy Fairbank and Dr Paul Pynsent of the Royal Orthopaedic Hospital, is currently being used by them to gain insight into the symptoms of back pain, and to try to find simple diagnostic tests that can be made in order to separate patients into meaningful categories. If this can be done, then firm diagnosis of the cause of the pain, when it is available, may allow a pathology or group of pathologies to be associated with each category.

The questionnaire is computer based, and is addressed to the patient rather than the practitioner (although a few variables such as height and weight are supplied by clinical staff).

The original intention of my project had been to use the replies, together with a clinical classification, as a training set to determine which questions would be most useful in discriminating between the classes using only the questionnaire. This turned out not to be feasible since the classification had not yet been made. It was also felt by the Birmingham team that it would be more useful to have an 'unbiased' (i.e. non-medical) assessment of the data, in view of the problems of the subjectivity mentioned earlier.

This report describes an attempt to analyse the replies to the questionnaire from a number of patients, to see if they fall into any 'natural' groups. The Birmingham team have been developing their own classification, based mostly on clinical judgement, and it is of interest to see whether their grouping can arise from the replies to the questionnaire.

The main statistical technique used was cluster analysis. This required a large amount of computing, which was done on the Cambridge University IBM mainframe. Programs were written either in Fortran or in Genstat.

2.3. Pottery and Politics

The aim of this project is to study decorative motifs on Iron Age vessels from the Argive Plain in Greece and to determine whether thay may be useful in explaining political changes that occured in the region.

A major problem is that very little is known about the history of the Plain; there are no contemporary sources (the art of writing was lost during the Iron Age) and the first written documents are not reliable. Hence our archaeological findings, and in particular pottery, are our main source of information. Therefore, we must continually be wary of using a circular argument to explain the vessels' significance.

The dataset consisted of stylistic information on 966 such vessels. It was very large, with sixteen variables being associated with the style of each pot. Originally, my task was to analyse it, in any way that I might choose. My one pointer was a paper by my own supervisor which employed the same dataset to explain political changes.

My interest lay in the painted motifs of the pots, so I decided to restrict attention to the variables concerned with decorative motifs, and to create a new variable, var 11, giving information about the overall painted style of a pot (rather than the painted style of a particular area of a pot). My initial intention was to cluster the pots and determine what motifs were responsible for the different clusters.

The clustering was done using cluster analysis. However, lack of true understanding of the very complex dataset led to clusters which were meaningless, at this stage. Hence, I decided that before trying any detailed analysis, it was first of all necessary to realise the potential of the (now restricted) dataset. This I did using SPSS-X.

The SPSS-X analysis demonstrated that, because of the way that the data had been collected, if any form of grouping was to be utilised, the easiest and most informative would be a grouping by site and period. Hence I decided to concentrate on the relationship between decorative motifs and the politics of the region. So, the project became concerned with the issues that my supervisor had considered. However, our analyses proceeded very differently.

The SSPS-X analysis also drew atention to the size of the site-period groups. The samples in all but the four principal sites were too small to be included in any analysis. Of the four remaining sites, namely Argos, Asine, Mycenae and Tiryns, great care had to be taken over some of the samples which were very small. However, I decided not to remove these samples as they could be useful in giving a better overall picture of the changes which occurred between the four sites over the four time-periods.

The analysis now involved studying such changes, and most of the project is concerned with these. Throughout, I concentrated on the decorative motifs of var 11, the variable that I had created, and I used var 6, the variable concerned with the decorative motifs of the main element of the main zone of a pot, as a check variable. From this stage, all computing was done using Genstat, a more sophisticated, versatile and powerful package than SPSS-X.

The analysis involved using a variety of techniques which included similarity matrices, Mann-Whitney tests, Multi-dimensional Scaling, and Correspondence Analysis. Of these, Correspondence Analysis was undoubtedly both the most powerful and the most instructive technique, and for that reason I have tried to explain how it works. It is also the technique which, I believe, would provide the best basis for any future work to be undertaken on this (or a similar) dataset.

2.4. Adult and Youth Unemployment; A Structural Analysis for the United Kingdom (Cambridge University Department of Applied Economics)

In this project I investigate youth and adult employment in the United Kingdom over the period 1974-1982. In particular, I considered the impact of overall strength of the economy, and the costs and productivity of youth and adult labour (both in absolute levels and relative to each other) on the development of employment in the two groups.

The originality in the analysis lies in its use of data which is disaggregated by industrial sectors, and some attention is paid to the choice of an appropriate level of disaggregation.

The results obtained are generally in line with those obtained by previous researchers using aggregate data, although there are significant differences between sectors. Data limitations mean that the results are to be treated with caution, but the work is suggestive that futher analysis would be worthwhile.

Extensive use was made of GLIM via the Statistical Laboratory Hewlett Packard workstations in order to carry out these analysis.

2.5. Two Models for the Distribution of Whales in the Antarctic and the Associated Abundance Estimates

For some decades, a systematic program of whale marking has been in operation in the North Atlantic, North Pacific and Antarctic. Numbered marks were fired into whales from marker ships, the marks being designed to bury themselves in the body of an animal and remain there without causing serious injury. Data on the recovery of marks during normal whaling operations was also recorded.

A considerable amount of work has been done on the data, in particular to estimate whale abundance using standard mark-recapture techniques.

This project considers the data principally for Minke Whales, but also for Fin Whales in the Antarctic, and assesses the fit of two formal models for the disposition of whales representing the following situations; firstly a 'diffusion' model, in which marked whales diffuse away from their marking position in a gradual fashion, as seen during successive summer seasons, and the distance a whale may have moved since marking increases with time; and secondly a 'home-range' model which imagines each whale to belong to one of several 'stocks', the distribution of any whale in a particular stock, in a given season, being geographically limited in some sense, and independent of the position of the whale in the previous season.

Population estimates in the spirit of each model are then suggested. For the diffusion model this consists of a series of Petersen estimates, where the overall number of marked whales in each region considered is estimated under the model. For the home-range model, abundance within areas assumed to contain the different stocks is estimated. Three methods are employed: the average of the Petersen estimates for each season, Chapman's multiple sample estimator (which is essentially a weighted average of Petersen's estimates), and a probability model for a single release followed by a continuous process of recoveries, in which mortality rate and population size are simultaneously estimated.

Programs in Fortran on the IBM 3084 Q mainframe were used to manipulate data, and statistical work was also done using Genstat on the mainframe. The project was prepared on an Apple Macintosh SE.

2.6. Analysis of Dolleman Ice Core Data (British Antarctic Survey)

In the 1950s it was found that an unusually high number of aeroplanes flying over the Antarctic crashed into the glaciers for no obvious reason. It was discovered that the planes' radar navigation systems showed them to be far from the ground even just before a crash. This led to investigation into the ability of ice to reflect the electromagnetic waves produced by the radar. The investigators concluded that the ice's ability to reflect the waves varied because of natural variations in its dielectric properties. The result of this variation was that the aircraft's radar signals were not always reflected by the ice on the surface but often by ice much deeper, giving the impression that the plane was much higher than it actually was.

Further investigation revealed that there was a marked difference in dielectric behaviour between ice removed form the cold interiors of polar regions and ice from warmer coastal areas. It was found that the ice from the coast had a large number of impurities that must have come from the sea, and it was postulated that the dielectric properties of the ice depended in some way on its chemical content. Thus in due course British Antarctic Survey collected data on ice cores for analysis.

The data consisted of several files each containing observations made at intervals along a 45m section of ice that had been removed from Dolleman Island in the Antarctic. There were 806 observations in each file. The observations can be divided into two categories, dielectric and chemical, depending on what they represent. The five chemical observations were combined into just two variables representing the salt and acid content. There were two dielectric variables σ_∞ and f_r.

The project consisted of investigating how the chemical variables could be used to model the dielectric ones and *vice versa*. All modelling was carried out using Genstat, on the University mainframe.

Chapters 2, 3 and 4 of the report explain what led BAS to the chemical and dielectric analyses and why they want to model the situation. Chapters 5 to 8 contain a description

of the data and the fitted models, including explanations of the theory behind them and the problems encountered in fitting them to the data. The statistical models were based on linear regression, but included an extension to raise the independent variable to a power. These models were fitted assuming that the random component comprised, firstly, independent Normal random variables and secondly, an autoregressive process of order 1. Some of the models had already been tried by British Antarctic Survey and so part of the project doubled as an independent check of their results.

Chapter 9 is a brief discussion of the computer programs that were written to carry out the analysis. The results of the analysis are given in chapter 10 and the conclusions that were drawn are in chapter 11.

2.7. Infant Mortality: Statistical Analysis of 10 English Parishes, 1581-1855

In this project I examine two statistical models for infant mortality in the past. I fit these models to data from 10 English parishes spanning the years 1581-1855. Both of these models consider mortality risk as a function of covariates which describe the birth circumstances of children. My aim is to construct a model which manages to capture some of the featues of infant mortality and which manages to show the relative importance of these features i.e. which circumstanes lead to higher rates of mortality.

The covariates I use are determined by what can be inferred from the raw data sources, which are almost exclusively ecclesiastical records of birth, deaths and marriages. This is mainly confined to the reproductive pattern of a mother, by which we mean such factors as the number, spacing and sex of the children born to a mother and her age at childbirth. I also consider the effect of geographical and seasonal factors on mortality.

The data I use has been collated by the Cambridge Group for the History of Population and Social Structure. Because of the wide variability in the quality of ecclesiastical registration a considerable part of my project involves devising programs to select a reliable subset of the complete data available and then to calculate the required information on childbirths from it.

The second part of the project involves using the statistical packages SPSS-X and BMDP to fit the models. The two models I consider are the parametric logistic regression model and the non-parametric proportional hazards model.

Both these models allow me to assess the importance of the various covariates in explaining the infant mortality process. I am particularly interested in sorting out the effect of the following factors.

(1) Age of mother.

(2) Interval since previous childbirth.

(3) Rank of childbirth within a family.

(4) Completed size of a family.

Among demographers, opinions differ on how these operate and which are most influential but my results seem to indicate that short intervals, first and higher birth rates and small completed family sizes are most correlated with high infant mortality.

2.8. Investigation of Ice-Floe Distribution and Orientation

This report is a collection of investigations into the properties of the sea-ice cover in Kong Oscars Fjord in East Greenland.

The **Preferred Orientation Distribution** for ice-floes was ascertained and found to be Normally distributed. The ice-floes, which are roughly ellipsoid in shape, align themselves with their maximum diameter at almost right angles to the length of the fjord.

The **Areal Distribution** of the floes was also considered and, through contacts with theory and careful attention to data quality issues, was found to obey a power-law relationship, formally known in statistical circles as the Pareto distribution.

An investigation of the shape of ice-floes led to a study of the **Area-Perimeter** relationship for the floes. The discovery of a non-fractal perimeter for the floes fitted in with the *a priori* supposition that the ice was well-established multi-year ice with a non-angular perimeter. The discovery also enabled the proposed study of the shape of the floes to proceed, using the shape computations from the digitization process.

The **shape** of the floes was generally ellipsoidal with varying degrees of elongation. The report also contains smaller investigations which provide pathwork for future work.

New methods in this area, such as the application of certain fractal ideas to ice, and maximum likelihood methods for parameter estimation were applied to these problems. A comparison of ice-areas was also an important new feature of experiments in this field.

The investigation required the use of various computing facilities (hopelessly unconnected) including Minitab, GLIM, Genstat and others on the University mainframe, Departmental minicomputer and various microcomputers.

2.9. Comparing Techniques to Analyse Interaction Effects in a 2×2 Completely Randomised Design

The aim of the project is to compare some methods for determining the interaction of two treatments. The type of data sets which one wants to analyse will typically be of non-balanced design, small sample sizes of unequal variance and non-Normal. Thus one will not be able to use standard methods, like analysis of variance. Although this project does not require a particular set of data, a set of results from a neuro-endrochronological experiment was provided as a typical data set of the type one wishes to analyse. The data given concerns the effects of two treatments on the release of some hormone from glands of rats.

The techniques which seemed most appropriate were randomisation and boot-strapping, both of which require no distributional assumption. These are compared with parametric tests which assume Normality, but are moderately robust to departures from this assumption. The bootstrap is a technique used to estimate the parameter of a distribution given a sample, using the sampling distribution. One can also calculate the standard error for the estimate. The randomisation test can be used to test a hypothesis about a distribution given a sample and a suitable test statistic, by generating 'equally likely data sets'. The parametric tests will be a t-test adapted to test for interaction, and the Satterthwaite version of the t-test, which allows for unequal variance.

To compare these tests one needs power curves. These were obtained by generating alternative hypotheses, analysing them using the above techniques, and observing how often the program correctly rejected the null hypothesis at various significance levels. Ideally this would be done by generating the alternatives from a large variety of distributions, but due to the amount of time the program takes to run, only the Normal distribution was used.

All the analysis was done using a Fortran program which accesses the NAG Fortran Library for generating random numbers, and calculating Student's t probability values.

2.10. An Analysis of Human Childhood Growth (Birth to Four Years of Age)

This project concerns the growth pattern of human infants between birth and four years of age. The aim is to describe a child's length, y, in terms of his age, t. An examination is carried out of the performance of a variety of models, each of the form

$$y = f(t) + \varepsilon$$

where f is the function defining the model (which may be linear or nonlinear in its parameters) and ε is an error term. Different individuals are characterised by different parameter values.

The data set used to test the models consists of two hundred and four Cambridge babies with varying numbers of measurements. The observations occur at set ages betwen birth and four years of age, but for many babies do not extend beyond the first year, or year and a half, of life. For each baby the models are fitted by linear or nonlinear regression, whichever is appropriate.

The ability of each of the models accurately to estimate the observed infant growth curve is examined. A fault common to many of them is a tendency cyclically to under- and over-estimate the observations. This suggests a failure to represent the growth curve. Four of the best models are analysed in detail, with emphasis placed on goodness of fit, behaviour of residuals and the distribution of the fitted parameters. One of the models is chosen as the optimum of those studied. The analyses also suggest the form of an extra term which could be added to each of the four models to improve the fit.

The use of mathematical models to describe childhood growth is discussed and it is concluded that such models have a valuable part to play in gaining knowledge of the growth process.

All of the computing involved in the project took place on the Cambridge University mainframe. The analyses of the data were performed using Genstat and high-quality graphs were drawn using the Camplot library routines accessed by simple Fortran programs. The project was written up using a combination of the mathematical text processing language TEX® and its derivative LaTEX®.

2.11. Analysis of Data From Quantitative Autoradiography

An experiment was carried out to determine the ability of grafts of embryonic neuronal tissue to restore the lesioned rat brain. The recovery was measured by the density of three different types of receptor and an enzyme density. A non-parametric technique, the Steel-Dwass procedure, was used initially, to look out for significant differences between density measurements. Some of the assumptions used for a linear model were looked at – the independence of errors from two areas within the foetal graft and the serial correlation of measurements through sections of the striatum. Linear models were first fitted to pairs of animals using Genstat programs. After looking at the fitted models individual models were fitted and the Tukey-Kramer procedure carried out to locate the significant differences. Some conclusions were drawn about the effectiveness of the treatment.

Endothelin levels were measured using quantitative autoradiography in various parts of human, pig and rat tissue. It was desired to isolate a subgroup with high Endothelin levels for further experimentation. Previously this had been done by dividing the tissue in three groups, the high, medium and low Endothelin groups and here non-hierarchical clustering was looked at as an alternative. The data were clustered using Genstat into various numbers of groups according to the partitions which minimized the within group sum of squares. Particular numbers of groups were found into which the data seemed to divide well. The clusters found were judged on various grounds as to how meaningful they were. A Fortran version of the Fisher algorithm was run to check if the Genstat method did in fact find optimal clusters.

Features of the Genstat 5 Language: 3

K I Trinder
NAG Ltd
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom *OX2 8DR*

1. Introduction

Here is another batch of items about aspects of the Genstat 5 language which are not necessarily obvious from the Reference Manual.

2. Use of the = Symbol

The = symbol has several uses and there is some potential for confusion. The following points might help.

(a) In options and parameters

The most common use of = is to separate option and parameter names from their settings. In this context, = is a punctuation symbol. For example:

```
READ [CHANNEL=3] STRUCTURES=Farm,Cows,Pigs,Sheep
```

If the option or parameter name is not given, then the = should also not be given. In the example, `STRUCTURES=` could be omitted and in fact usually would be omitted as this is the first parameter.

(b) For assignment in expressions

The = symbol is used to assign the values that are generated in a (sub-) expression to an appropriate data structure. The most likely occurrence of this will be in a `CALCULATE` statement; eg:

```
CALCULATE C = A+B
```

In this example, there is no parameter name and therefore no = separator. In this case, the = is an assignment operator. Note that assignments may be embedded within expressions and that expressions may contain lists of assignments. For example:

```
CALCULATE C = A+B
CALCULATE D = C+B
```

may be given as:

```
CALCULATE D = (C = A+B)+B
```

or: `CALCULATE C,D = A,C+B`

although not:

```
CALCULATE D,C = C,A+B
```

since the order that the items are given in the lists is important.

Both of these shorthand techniques can be useful in improving the efficiency of a Genstat job or procedure, although it is harder to follow the sequence of operations that make up the whole expression.

A situation to be wary of is where there is ambiguity in the interpretation of =. Suppose the expressions `N=NVALUES(X)`, `M=MEAN(X)` and `V=VARIANCE(X)` are to be respectively stored in the structures `E1`, `E2` and `E3`. The statements

```
EXPRESSION [N=NVALUES(X)] E1
EXPRESSION [M=MEAN(X)] E2
EXPRESSION [V=VARIANCE(X)] E3
```

appear to be correct (assuming that the VALUE option name can be omitted), but in fact none are because in each case the letter before the = will be assumed to be an (abbreviated) option name. The first statement will fail because N will not be recognised as a valid option name. The second statement will fail because M will be recognised as an abbreviation for MODIFY, and MEAN(X) will not be a valid option setting. The third statement will work but V will be recognised as an abbreviation for VALUE, and the wrong expression, VARIANCE(X), will be stored in E3. The problem is easily avoided by either putting the expression in brackets, as in:

```
EXPRESSION [(N=NVALUES(X))] E1
```

or by giving the option name explicitly:

```
EXPRESSION [VALUE=M=MEAN(X)] E2
```

or by using the VALUE parameter:

```
EXPRESSION E3; VALUE=!e(V=VARIANCE(X))
```

(c) For testing equality in relational expressions

The relational operator for testing equality is ==. For example, in:

```
IF Alpha==0
  statements
ENDIF
```

assuming that Alpha is a scalar, *statements* will be executed only if the logical expression Alpha==0 is *true*; that is, if the value of Alpha is zero. Care should be taken when using this operator because if instead:

```
IF Alpha=0
  statements
ENDIF
```

had accidently been given, then Alpha would be assigned the value zero, the logical expression would always be *false* and *statements* would never be executed.

Note that a synonym for == is .EQ. and it is perhaps safer to use this. Also, it should be remembered that there is a different relational operator, .EQS., for testing string equality.

(d) In other compound relational operators

The symbol is also used in the relational operators /= (for non-equality), <= (for less than or equals) and >= (for greater than or equals). /= has the synonyms <> and .NE., and <= and >= have the synonyms .LE. and .GE. respectively.

3. Using Missing Values

The two following sections show how one technique with missing values can be used in quite different ways. The technique involves replacing some of the items of a structure with missing values by dividing those items by zero. Such a calculation normally generates a warning and this can be suppressed by using the JOB or SET directive with the DIAGNOSTIC option set to just errors. Note that care should be taken if the structures to be operated on, (Group and Count in these examples), already contain missing values.

(a) Printing numeric values as spaces

It is sometimes convenient to print numeric data structures, such as tables, with items of a particular value such as zero not being displayed; that is, spaces are to be printed instead of zeroes. Assuming that the structure is Count and that Count has no missing values beforehand, the following three lines will enable this:

```
CALCULATE Count = Count/(Count.NE.0)
PRINT [MISSING=' '] Count
CALCULATE Count = MVREPLACE(Count; 0)
```

If Count did have missing values beforehand, then the same thing could be done using a temporary structure:

```
CALCULATE Temp = Count/(Count.NE.0)
PRINT [MISSING=' '] Temp
DELETE Temp
```

In this case all of the items of Count with zero or missing values would be printed as spaces.

Note that in Release 2, the first step of this operation can be done in a more obvious way using the new function MVINSERT:

```
CALCULATE Count = MVINSERT(Count; Count.EQ.0)
```

This removes the need for a division by zero and there will be no consequent warning message.

(b) Omitting factor levels in regression

Suppose a regression analysis is required where one of the terms is a factor and the units associated with one level of the factor are to be excluded. An obvious way to achieve this is by using the RESTRICT directive, as in:

```
UNITS [15]
FACTOR [LEVELS=3; VALUES=5(1...3)] Group
READ Ep,Rp
data :
RESTRICT Rp,Ep,Group; CONDITION=(Group.EQ.1).OR.(Group.EQ.2)
MODEL Rp
TERMS Group*Ep
FIT Group*Ep
```

Genstat will readily perform the analysis, but will attempt to fit the third level of Group even though there are no units for that level after the restriction. Warning messages will be given and the parameter list will include the parameters associated with the third level of Group. To avoid this, Group must be redeclared with two levels and this can be done by replacing

```
RESTRICT Rp,Ep,Group; CONDITION=(Group.EQ.1).OR.(Group.EQ.2)
```

with

```
CALCULATE Group=Group/(Group.NE.3)
FACTOR [LEVELS=2; VALUES=#Group] Group
```

The job should then work as desired. To restore Group to its previous state, insert the following lines after the analysis:

```
FACTOR [LEVELS=3; VALUES=#Group] Group
CALCULATE Group = MVREPLACE(Group;3)
```

4. Bits and Pieces

(a) Wide data records

When Genstat is reading data, it does so up to the currently defined maximum width of record for the input channel. If you do not specify this maximum explicitly (in an OPEN statement), the default maximum of 80 characters is assumed. Characters beyond the limit in any record will be ignored: this can cause unexpected results if you do not allow for it.

A data file might contain the following, where the records are actually 90 characters wide:

```
467     376     736     235     234     345     158     893     188     782
937     179     790     162     949     332     527     267     723     731
:
```

The following job listing illustrates the potential problems:

```
1  OPEN 'filename'; CHANNEL=2
2  READ [PRINT=data,errors,summary; CHANNEL=2; SETNVALUES=yes] X

 1        467      376      736      235      234      345      158      893       18
 2        937      179      790      162      949      332      527      267       72
 3  :

 Identifier   Minimum      Mean   Maximum    Values   Missing
          X      18.0     426.5     949.0        18         0

3  PRINT [WIDTH=88; ORIENTATION=across] X; FIELDWIDTH=8

        X   467.0   376.0   736.0   235.0   234.0   345.0   158.0   893.0   18.0

        X   937.0   179.0   790.0   162.0   949.0   332.0   527.0   267.0   72.0
```

The ninth value on each line has been read incorrectly and the two tenth values have been ignored altogether, thus making X of length 18 instead of the intended 20. Note that in this situation it is impossible for Genstat to detect these errors, and therefore no errors are reported. You should be wary of this type of error, which could easily go unnoticed. The problem is corrected by using the WIDTH parameter of OPEN and the first line in the example should have been something like:

```
OPEN 'filename'; CHANNEL=2; WIDTH=90
```

(b) Using the repetition symbol (&)

A Genstat directive name which is to be repeated several times consecutively can be replaced in the second and subsequent statements by the ampersand character, &. For example,

```
HCLUSTER [PRINT=amalgamations; METHOD=singlelink] Smat
& [METHOD=completelink] Smat
```

will perform two cluster analyses using different methods.

There is one important rule to remember when using the repetition symbol: any option settings in the previous statement are carried over unless they are explicitly reset. Thus the setting of PRINT in the first statement of the example will also apply to the second clustering. One place where it is easy to overlook this rule is in the declaration of factors. The statements:

```
FACTOR [LEVELS=!(11,12,13,14,15)] Age
& [LABELS=!t(male,female)] Sex
```

will give an error after the second statement because the two labels do not correspond with the five levels carried over from the previous FACTOR statement.

Procedures may also be repeated using the repetition symbol although you should be aware that in Release 1 the option settings are not carried over. This will change in Release 2 and procedures will behave as directives when they are repeated.

(c) Extracting margins from tables

The following simple example illustrates how the values in the margin of a table, Tab, may be extracted and placed in a variate, Tabmeans. Tab is assumed to be classified by several factors, one of which is Fx. Tabmeans is to contain the means of the values of Tab over the classifying factors other than Fx. The temporary table Mvals may be deleted subsequently.

```
TABLE [CLASSIFICATION=Fx] Mvals
CALCULATE Mvals = TMEANS(Tab)
VARIATE Tabmeans; VALUES=Mvals
```