GENSTAT NEWSLETTER

NO. 2

JUNE 1976

Apart from two contributions, one from East Malling and one from Welsh Plant Breeding, this newsletter has been written by staff at Rothamsted. If you think you have no useful tips to pass on an article about experience with Genstat at your institute would be of interest. If you have praise it may inspire programmers to higher things; if you have blame it may persuade them to mend their ways. So let us have your contributions. We shall be pleased to print them.

Amendments and new pages for the manual and user's guides are attached.

Genstat at East Malling Research Station

Most of the statistical analyses of data at East Malling are channelled through the Computing Unit via our statisticians, and consequently the bulk of the Genstat programming is handled by two people. We felt, though, that Genstat's potential was not being exploited to the full because of a communication problem. In the Computing Unit we are aware of the facilities available but not always of the experimenter's requirements. Conversely if experimenters are not aware of what can be done they may refrain from asking for it.

To try to overcome some of this difficulty, and to get experimenters themselves involved, two short courses were held here. The aim was not to give the kind of comprehensive introduction which the Rothamsted course provides (many of our customers felt they could not justify spending 3 days learning the language), but to give a sufficiently thorough grounding in the directives which in our experience were most used, so that people would be able to run simple jobs, and would see the possibilities of Genstat for their own applications.

The first course of two afternoon sessions in June was a review of the kind of facilities Genstat could provide with only a brief explanation of the detailed programming involved. The second course of 3 afternoon sessions spread over two weeks in November was restricted to people who had some programming experience, or who had at least used a terminal. Eight people had no previous experience of Genstat and of these four have since become users of the system. Instruction was given in the use of data-handling directives, regression, simple analysis of variance and graph. A brief description of the directives covered and their syntax was prepared as a hand-out for the course. It was felt that after such an introduction people would be willing to refer to the Genstat Manual for additional facilities, and this has in fact been the case.

Most of the Genstat work handled by the Computing Unit is of the regression or analysis of variance type. Some fairly complex designs are involved, and also sheer number of data is sometimes a problem with some of the overseas trials which are analysed here on behalf of the Ministry of Overseas Development. Graphing of regression lines, and plotting of fitted values against residuals saved from ANOVA is increasingly popular, and PCP and CVA are used occasionally. In dealing with overseas data aberrant values can be a problem because of the remoteness of the experiment and experimenter, and because of the aforementioned problem of number of data, and for this reason residuals from ANOVA's of overseas data are fed into a macro which draws histograms of their distribution, which enables aberrant values to be picked out easily.

One application for Genstat which is sometimes overlooked because of the large and complex nature of the program is that of a data summary program. Often data summarisation/re-organisation tasks are so simple in concept that one-off FORTRAN programs are contemplated for particular tasks, but with its flexible data-handling facilities Genstat can often be used instead - dispensing with most of the problems incurred when developing a new program. We have used Genstat to knock data into shape and output them to DSET8 for re-input to another program with less flexible input routines. The only things to remember are the line-length and the need to run the Genstat DSET8 through DELIC to remove the carriage-control characters before re-input. (see note on secondary output channel in Newsletter No. 1 and note on links with other programs in this newsletter).

To give an idea of the volume of work involved - the Computing Unit's throughput during the year October 1974 to September 1975 was:-

4401 analyses of variance
1331 analyses of covariance
241 regressions
5 multivariate analyses

} of which about 90% would have been done on Genstat

and 860 other jobs of types not suitable for processing by Genstat.

Carole Pilcher
E.M.R.S.

Links with other programs.

There are built in links with two other programs in Genstat; RGSP [GRM 12.1] and SYMAP [GRM 12.2] . It is often useful to link Genstat to other programs (e.g. MLP or GLIM) and this can usually be done without the use of specially written interfaces.

'READ' can accept data in many different formats although they must be from character (line) files (e.g. those produced by FORTRAN formatted WRITE statements) so results from many other programs can be read by Genstat.

To produce results from Genstat suitable as input to other programs the following method can be used. Before printing the results to be transferred switch to the secondary output channel (see Newsletter No. 1). Then set up CAPTIONS and/or HEADINGS to hold instructions and appropriate data terminators for the other program. Then PRINT the Genstat results suppressing all labelling and print the CAPTIONS holding data terminators and instructions where necessary. The file created on the secondary output channel can be used as input to the other program (after using DELIC on the 4-70).

Example:

Genstat has created three variates of counts and of totals for a number of groups. These are to be transferred to MLP for Probit Analysis. The Genstat code is

```
'OUTPUT' 2 $ 80
'SCAL' L
'CAPT' ' ' CAPTION  **** Counts and Totals for grouped data **** ;
           DATA (275, 3) ' '
'FOR' DUM1 = 3(GRP); DUM2 = CNT(1...3); DUM3 = TOT(1...3)
'PRIN/S, LABR = 1, VAR = 1'  DUM1, DUM2, DUM3 $ 6.0
'JUMP' L * (DUM2  .IS.  CNT(3))
'CAPT' ' '  /  ' '
'REPE'
'LABEL' L
'CAPT' ' '  ;
          PARAM 1 0 0 1 1 1 ;
          FIT PROBITS ;
          END ' '
'OUTPUT' 1
'RUN'
```

N.G. Alvey
R.E.S.

Setting values missing

To set values missing if a condition is false write

'CALC' Y = Y + 0/(Condition)

Fred Potter
WPBS

Termination Codes

(System 4 versions only)

Genstat 3.08 will issue termination codes which should allow sensible sequencing of (Multijob) Jobs. Since only one termination code may be given for a Job, it can only be an overall assessment of a batch of (Genstat) jobs; in critical cases it may be desirable to submit jobs one at a time and not batch them.

The possible values of the termination code, and their meanings, are:-

Hex	Decimal	
00	0	An error has occurred that is not trapped by Genstat (usually an I/O error) and the program has not in fact issued a termination code.
01	1	Normal ending, no diagnostics
OC	12	Normal ending, but at least one diagnostic comment has been given
FC	252	At least one job in the batch has failed with a fatal diagnostic

Given that a Job will issue a termination code tc, subsequent Job may be prevented from running if

(a) tc $\neq 1$

or (b) tc does not lie in the range $[0, 12]$
depending on how particular you wish to be.

Text Indicator

After suffering an unfortunate experience, a Genstat user suggested that the Genstat compiler should indicate which lines of input it regards as part of a text. Genstat 3.08 will do this, by preceding the line number by a minus sign when reading text.

More specifically (if you are interested), when an initial prime pair '' is read an indicator is switched on; when the next prime pair is read the indicator is switched off. When a line is printed - which happens after all of it has been scanned - then if the indicator is on the line number will be negative.

Howard Simpson
R.E.S.

Restricting the unrestrictable

Most Genstat directives now accept restriction. READ does not, and is not likely to - after all RESTRICT was introduced to allow analysis of subsets of data and a restrictable READ would have no more effect (and be less flexible) than a normal READ followed by a RESTRICT - though it might be useful by allowing subsets of a large body of data to be read).

There are however some directives which do not currently allow restriction when it might be necessary (SYMAP, for example). The following sets of instructions will solve the problem:-

```
'RESTRICT'  VARIATES $ F = - - -

'OUTPUT' 2

'PRINT/P, LABR = 1 ' VARIATES $ - - -

'CAPTION'  ''  'EOD' ''

                  and

'INPUT' 2

'READ/P, NUN = V ' VARIATES
```

Note:

(1) The second set of instructions must be run as a separate job - System 4 at least will not allow the output file (DSET8) to be used as an input file (LFILE2) in the same job.

(2) The field widths in the PRINT command must be large enough to print all values in F - format with separating spaces.

(3) In the CAPTION the space after 'EOD' is essential.

Howard Simpson
R.E.S.

Macros for regression in tables (TREG and SREG)

It is sometimes desirable to study the relationships of regressions to factors used for indexing the data. Results printed as tables clarify the picture. When the regressions are calculated by operating on tables with margins, the marginal values correctly summarise regression within the table. Thus the cell for the grand mean contains the regression that ignores all indexing factors. An alternative option of pooled within-cell regressions in the margins can be obtained by setting the scalar PR = 1.

Two MACROS are provided, TREG operates on vectors of one _y_ variate and one or two _x_'s, together with the indexing factor vectors. SREG operates on tables of variates and their sums of squares and products. If there are two *x*'s the program calculates both simple and partial regressions. All regressions are with intercept. A listing of the macros can be obtained from the author.

Input Minimum setting of 'REFERENCE' options is NID = 50, NUNN = 20.

1. TREG: The user's program must contain the following directives:

   'SCALAR' NX = no. of *x*-variates (1 or 2) [: PR = 1]

   'SETS' EXES = *x*-variate identifiers(s)

   : WYE = *y*-variate identifier

   : CSET = classifying set for tables

   'USE' TREG \$

   'SAVE' Tables of results (or 'PRINT' if suitable print digits can be specified).

2. SREG: Tables of variate sums and their SSP must have the identifiers shown below. Single length operations are used, so the user must subtract a constant from each variate when forming the tables and supply the list of constants in the C scalars. The notation below is in terms of deviations from these constants. The user must ensure that if a value for one variate is missing, the other variates from that unit are set to the missing value. *n* is the associated count table.

   (a) one *x*

   T(1..6) = x^2, xy, y^2, x, y, n

   (b) two *x*'s

   T(1...10) = x_1^2, $x_1 x_2$, $x_1 y$, x_2^2, $x_2 y$, y^2, $x_1 x_2$., y, n

   'SCALAR' NX = no. of *x*-variates: C(1, 2) or C(1...3) = constants subtracted from x, y or x_1, x_2, y respectively [: PR = 1]

   'SETS' CSET = classifying set for tables.

   'USE' SREG \$

   'SAVE' or 'PRINT' results

Results. TREG and SREG both produce the same results.

1. Univariate regression

T(4...10) contain respectively $\bar{x}$, $\bar{y}$, n, regression coefficients (b), intercepts, S.E(b), and % variance accounted for.

2. Bivariate regression

a. Partial regressions:

$T(15, 16) = b_1, b_2$

$T(17)$ = intercepts

$T(18, 19) = S.E.\ (b_1, b_2)$

$T(20)$ = % variance accounted for

b. Simple regressions

$T(7...10) = b_1\ a_1,\ SE(b_1)$, % variance accounted for

$T(11...\ 14) = b_2,\ a_2,\ SE(b_2)$,% variance accounted for

The macros do not print results. The user is recommended to save the results in backing store because prior specifications of print parameters may turn out to be inappropriate.

F.B. Leech
R.E.S.

New Regression Facilities

At present, the regression section of Genstat is being altered in order to allow more general analyses to be carried out. While every effort is being made not to change existing conventions, it is inevitable that some alterations have to be made. It seems best at this stage, while the coding is still being done, to set out the proposed alterations so that users can comment on them.

There will be two major extensions to the present facilities. Firstly, model formulae may be used, both in SSP matrices and in regression operations. This will allow interaction terms to be fitted in a regression model in the same way as main effects are now fitted, and will be particularly useful for analysing unbalanced experiments which are not suitable for the 'ANOVA' directive.

The second extension is to allow the use of 'generalised linear models' (GLM). The conventional regression model is a particular instance of a GLM; so are Probit analysis, log-linear models for contingency tables and analysis of variance components.

SSP matrices

All directives involving SSP matrices are acceptable to the new version, and will have the same effect. However, the matrices that are produced will differ in one way from the present ones - there will be no row corresponding to the last (intrinsically aliased) level of a factor. This will only be noticed if matrices are to be read in direct, or printed. The intention of this change is to reduce the size of the matrix when analysing, for instance, an unbalanced design with several factors of two levels. However, if any users have a particular need for explicit representation of all levels, it might be feasible to add an option to the 'DSSP' directive to allow this.

Y- variates

At present, any number of variates in the regression model can be specified as y-variates at one time, except for the directives Best, Worst and Minimise. This will remain true for the standard regression model, but in the other GLM's, involving iterative methods, only one y-variate will be recognised. In future, it will not be possible to declare a variate to be a y-variate if it is already fitted as an independent variate: it must be 'DROPped' from the model before the new 'y'-directive.

Regress directive

At present, when an SSP matrix is calculated, either by an 'SSP' or by a 'REGRESS' directive, a subset of units is used, depending on previous 'RESTRICT' directives and the presence of missing values in the variates or factors. In future, unitz of data will also be excluded if there is a missing value for the weight variate or the grouping factor, if these facilities are being used.

X-set directives

There should be no change in the results of these directives, except that the default regression coefficient for the last level of a factor will no longer be shown. In addition to the extra printing and saving that will be allowed for iterative models, it is proposed to allow a summary analysis of variance table to be compiled for a sequence of directives. This will show the changes in residual sums of squares, degrees of freedom and the terms that have been added or dropped.

Any comments or suggestion concerning these changes would be welcome.

Peter Lane
R.E.S.

Genstat Reference Manual

Amendment List No. 8 (for April 1973 Manual)
Amendment List No. 4 (for January 1975 Manual)

Note: *after line number indicates lines counted from bottom of page
±indicates line number for January 1975 manual and Ø for April 1973 manual if the line numbers do not coincide with each other.

Page No.	Line No.		Amendment
Contents p3	7 Ø	8±	After line 7 Ø, 8± line should read "2.2.1.1 Option of USE"
	1*		After line 1* add "2.6.1 Scope of special structures"
Contents p15	6,7,8		Delete lines 6, 7 and 8.
1.1 p1	19		After line 19 delete sentence beginning "See [A2]"
1.3 p2	3 Ø		After "identifier!" add "The ! distinguishes premultiplied lists from subscripted identifiers"
	12 Ø	13±	After line 12 Ø, 13± add "Brackets can be used to eliminate ambiguities".
6.5 p3	8		Alter "ASSOCIATED COUNT" to "ASS.CT."
	14		Put full stop after "position". Delete rest of line. After line 14 add "Identifier. If this is a structure with"
7.2 p1	1*		After "they left off " add "The workspace structure is a special structure [2.6.1]."
8.1 p1	9*		Line should read "be restored by a directive with null argument. These three directives define special structures [2.6.1]."
8.9 p1	14*		Alter "means (not extracted" to "means (treatment terms only, not extracted"
10.6 p1	2*		Line should read "C Print squared centroid-distances"
A1 p2	13		After line 13 add "CA-16 Matrix singular"
A3 p1	19		Line should read "EFF = R, SE = D, ORTH = N"
A3 p6	9		Line should read "'RESTRICT [/option]' ilist [Ø restrictors [= restricting sets[; integer vector]]]] 6.7.3"
A4 p1	1±		Before line 1± add "APPENDIX 4"
	6 Ø	5±	After line 6 Ø, 5± add "ABS"
	11 Ø	10±	After line 11 Ø, 10± add "DET"
Index p1	5		After line 5 add "ABS ( ) 6.2.1"
Index p2	16 Ø	19±	After line 16 Ø, 19± add "DET ( ) 6.3.1"
Index p3	32		Line should read "INPUT ( ) 6.2.1/6.6.1.1"