GENSTAT NEWSLETTER

NO. 4

OCTOBER 1977

The new manual is now being distributed and some of you will be receiving your copy with this newsletter. We hope you will find the new manual more helpful than the old but we will be grateful to hear of any errors or obscurities that still exist. These can be covered in future updates.

The promulgation of Genstat would have been impossible without those who convert and distribute the program. We would like to take this opportunity of thanking the following for their very helpful collaboration:-

Dr. Russ Cormier	Statistical Research Service, Ottawa
Dr. Peter Baghurst	Waite Institute, Adelaide
Dr. Paul Griffiths	Oxford University
Dr. Errol Jones	Cornell University
Dr. Yien Kwok	Manchester University
Mr. Nick McLaren	Cambridge University
Mr. George Paterson	Shell Research
Dr. Julian Read	Program Library Unit, Edinburgh
Numerical Algorithms Group	Oxford

The help of future convertors and distributors will be acknowledged in subsequent newsletters.

OPTIMISATION IN GENSTAT

Optimisation facilities are being introduced in 4.01 on a trial basis. The relevant sections of the manual have not yet been written, and the following is a description of the facilities currently available.

Optimisation requires two directives, a 'MODEL' directive which is a sequence of CALCulate statements defining the expected values or working regression variates as a function of parameters, and an 'OPTIMIZE' directive. The calculations defined by 'MODEL' are carried out under the control of 'OPTIMIZE'.

The 'OPTIMIZE' directive computes a sum of squares (or log likelihood or a general function) for trial values of parameters either as part of a search for a minimum or on a systematic grid which can be subsequently displayed and linked to the 'GRAPH' or 'CONTOUR' directives.

Description of directives

There are two directives.

'MODEL' Name $ sequence of expressions where expressions are arithmetic or logical expressions separated by $.

Examples 'MODEL' EXPM $ Z = A + B * R ** X

'MODEL' BLIN $ ZI = (X .GE. C) * (X - C)

'OPTIMISE / option list' sequence of nameable lists
where option list is

ITER, LIK, NPAR, CONST, GRID = integer, PRIN = letters

and nameable lists are

MODEL = pointer;	(a name of a previous 'MODEL')
FMIN = scalar;	(the minimum value of the log likelihood when formed)
PARAM = list of scalars;	(parameters named in the model, including linear parameters)
STEPS = variate;	(step lengths for trial differences)
UPPER = variate;	(optional upper bounds for parameters)
LOWER = variate;	(optional lower bounds for parameters)
Y = list of variates;	(observed data)
Z = list of variates;	(expectations computed by 'MODEL')
WT = variate;	(weights for least squares or binomial errors, 1 by default)
VCOV = symmat;	(the computed dispersion matrix of parameters)
RES = variate;	(optional stored weighted residuals)
FVAL = variate;	(fitted values when different from Z)
DERIVS = list of variates;	(approximate derivations of fitted values)
ARRAY = variate;	(output for grid evaluation)
EXIT = scalar;	(indicator of successful convergence).

OPTIMISE sets up a likelihood function linking the working variates Z computed by MODEL with the observation variates Y, according to the method specified by LIK. The likelihood is either a direct function of NPAR scalar parameters PARAM or is a conditional likelihood for NPAR non-linear parameters in a model involving linear parameters. Observations may be weighted by variate WT.

If GRID = 0 the likelihood is optimised from starting values PARAM using initial steps STEPS, optional constraints LOWER and UPPER, and the resulting solution in FMIN with parameter estimates in PARAM and dispersion matrix in VCOV. EXIT is 0 if convergence is satisfactory, 1 if more than ITER iterations have been used, 2 if bounds have been exceeded.

Fitted values may be output to FVAL, residuals to RES and first derivatives of FVAL with respect to each parameter in DERIVS.

If GRID > 0 the function is evaluated at each grid point within the NPAR-dimensional array from LOWER to UPPER of each parameter divided into GRID-1 intervals, giving rise to GRID ** NPAR values stored in ARRAY.

Options

(1) ITER = 20 If number of iterations exceeds this, EXIT = 2.

(2) LIK = 1 Likelihood, or any other function with a minimum, in FMIN.

2 Normal, weighted sums of squares of differences of Y and Z (for log normal take logs of Y and Z).

3 Normal, where the expectation is a linear function of Z variates which involve non linear parameters. Linear parameters are automatically placed in the (NPAR + 1)th parameter onwards, the Constant term last.

4 Poisson variates with expectation Z.

5 Poisson variates with expectation N * Z, where N is an extra linear parameter.

6 Binomial variates with observed proportion Y of total WT, and expected proportion Z.

7 Multinomial distribution with counts Y and proportions Z that add up to 1.

(3) NPAR Number of parameters to be estimated ($\leq$ 6). If LIK = 3 or LIK = 5 it is the number of non linear parameters, which becomes extended automatically to the full set of parameters on output, depending on the number of working Z variates and the value of CONST below.

(4) CONST = 0 (1) to include (exclude) the constant term in the model when LIK = 3.

(5) GRID = 0 Optimise likelihood.
2 or more, compute GRID ** NPAR values in ARRAY.

(6) PRIN = letters to control printing, which may be

P Parameters, standard errors, correlations, and residual mean square or chi squared.
M Convergence monitoring.

Method of use

1. General considerations

The optimisation facilities are intended for use with well-conditioned parameterisations of models with few parameters. They are not intended for problems such as maximum likelihood factor analysis, or for large non linear factorial effect models for which the generalised linear model facilities of 'REGRESS' should be used.

A well-conditioned parameterisation is one for which

(a) no parameter can possibly take infinite value

(b) no two parameters are highly correlated because they essentially estimate the same quantity

(c) the likelihood is not too asymmetric at the optimum with respect to any combination of parameters.

Neglect of these principles is the cause of much of the difficulty associated with optimisation. The conventional parameterisations of most non linear models are nearly always ill-conditioned for some data sets, and therefore if optimisation is to be used efficiently on a routine basis the parameterisation must always refer to working constraints derived from the data (such as empirical means, ranges, orthogonal polynomials, sample variances). The conventional parameters may be calculated from the working stable parameters once the solution is obtained.

A solution is not guaranteed as (1) it may not exist if the model does not permit any form of the expected value variate that resembles the data more exactly than a limiting sub-model, (2) it may not be unique if the data are insufficiently distributed over the possible range, or if the model is essentially ambiguous because totally different parameter values generate the same response.

2. Suggested settings

If the parameters are expected to lie within a finite range then suitable values for STEPS are in the order of one per cent of the anticipated ranges. If steps are too small numerical differencing (considering that 'MODEL' may only be to single precision) may break down, whereas if they are too large gradients may be unreliable and convergence premature. Convergence is tested by the relationship of final adjustments to step lengths.

For curve fitting and allied least squares problems the form LIK = 3 should always be used; being safer, faster and simpler to specify and to study.

ITER should not be increased if the model fails to converge. Instead the data should be checked or the model re-parameterised.

GRID should be used carefully, otherwise an excessive amount of time will be used or output generated.

The bounds UPPER and LOWER should be used if some parameter values would cause computing errors such as log of negative argument or exponential overflow. More general constraints must be included as logical functions within the 'MODEL' for example by computing an extra term (CONSTR .GT. 0) * K * CONSTR which imposes a penalty on exceeding the constraint which can be controlled by setting different values of K.

3. Example

To fit the exponential curve with normal errors,

$$E(y) = a + br^X$$

it is best to note that a and b are linear parameters. The estimate r depends on the scaling of X, and it is necessary to ensure that r is somewhere in the interval (0.02, 0.98) otherwise values too close to 0 or 1 make it difficult to estimate a and b. a and b are very unstable as r approaches 1 (a straight line) and tend to infinity in opposite directions. The form exp(-kx) is also unsuitable because k can go to infinity if the first data value of y is very different from the remaining data values.

Therefore first scale X, say by its standard deviation. Then writing

```
'SCALAR'  R, B, A
'VARIATE'  STEP, MAXR, MINR $ 1
'MODEL'  EXPM $ Z = R ** X
'VALUES'  STEP = .01   :  MAXR = .98   :  MINR = .02
'OPTIMISE / LIK = 3, NPAR = 1'  EXPM ;  FMIN ;  R, B, A ;  STEP ;  MAXR ;
  MINR ;  Y ;  Z ;  ;  ;  ;  FVAL
```

should lead to a rapid solution, unless the data curves the wrong way, in which case reversing the sign of X will give a fit.

Intending users should consult the following papers:

Ross, G.J.S. (1970). The efficient use of function minimisation in non linear maximum likelihood estimation. Appl. Stats. 19, 205-221.

Ross, G.J.S. (1975). Simple non linear modelling for the general user. Proc. 40th Session Int. Stat. Inst. Warsaw. 2, 585-593.

Note that there is no intention to include within GENSTAT the special routines in the Maximum Likelihood Program (MLP) which give special consideration to a particular range of models.

New facilities will be provided in later releases, and the present version, though tested, is not likely to be foolproof, especially if the model is inconsistent or the data unsuitable.

Gavin Ross
R. E. S.

NEW LIBRARY MACROS

Brief descriptions of macros that have been added to the library recently and that were not described in User Guide No. 9 follow. Full descriptions of library macros are not given in Chapter 10 of the new manual but descriptions of particular macros can be obtained from the Programs Secretary.

Multivariate analysis of variance

A macro MANOVA for multivariate analysis of variance is now available. The user must supply a list of variates for analysis with the appropriate treatment and block formulae for the design employed. The error and hypothesis SSP-matrices will be calculated and printed as well as several test criteria, including Wilk's Lambda and Roy's Maximum Root. The canonical variates for each error stratum are also produced. The macro cannot, however, produce a multivariate analysis of covariance.

Starting classification

A macro CLASSF to obtain a starting classification for use with the CLASSIFY directive is also available. Given n units this will find the k units (k < n) that are furthest apart and use these as the nucleii for k classes. The remaining (n - k) units will then each be allocated to the class having nearest nucleus.

Colin Banfield
R. E. S.

Generalized linear models

The macro GLM is no longer available in the macro library, since generalized linear models can now be fitted using the regression directives. However, a new macro called GLMODEL has been provided, to allow models with non-standard link functions and/or error distributions to be fitted. The user must provide short macros to perform such calculations as formation of the fitted values from the linear predictor, or calculation of the log-likelihood. Other facilities in the macro are the same as were provided in GLM.

Probit analysis

It is now possible to fit a range of models to quantal data from bioassay, using either probit or logit link functions. The estimated parameters for these models are not however the quantities which are usually required, i.e. estimates of LD50, LD90 etc. The macro FIELLER requires as input the coefficients and their variance-covariance matrix, as produced by the regression directives, and will calculate estimates of specified log-dose percentages with standard errors and 95% fiducial limits. The estimates can be produced in one of two forms: actual estimates of log potency, or potencies of treatments relative to a standard.

Censored data

Data values from an experiment are said to be 'censored' if their actual values are not known, but it is known that they are greater than (or less than) some limit. This can occur for example if the observations are times to occurrence of some event, and the experiment is concluded before the event has occurred for some experimental units. The macro CENSOR can be used as a pre-processor for censored data from an experiment which is analysable by the ANOVA directive. It produces estimates for the censored points, and a value for the effective number of residual df - which can be used to correct the standard output produced by ANOVA when the processed data are analysed.

Peter Lane
R. E. S.

Aliased model terms in analysis of variance

Any aliased model terms in an analysis of variance are listed in the information summary, printed below the analysis-of-variance table. In versions of Genstat before 3.06 the terms with which each model term was aliased were also given. In 3.06, however, a new dummy analysis method was introduced to allow orthogonal designs to be more efficiently analysed and to give better numerical accuracy. Unfortunately, the new method does not allow the aliasing relationships to be readily determined. In some designs it may be clear with which terms a model term is aliased; if this is not the case, macro ALIAS can be used. This performs an analysis of variance on a variate specially generated for the model term concerned. Terms with which the model term is aliased are indicated by non-zero sums of squares in the AOV table printed by the macro.

Roger Payne
R. E. S.

GENSTAT MANUAL

Amendment list no. 1

Note: * after line number indicates lines counted from bottom of page.

Line count does not include lines with page or release number.

Page no.	Line no.	Amendment
Part I		
Ch. 2 2p3	20	Alter "symmetrix" to "symmetric"
Ch. 3 4p1	17*	Alter "This function" to "The function ELEM"
Ch. 4 4p1	2*	Delete "at least"
Ch. 4 8p3	15	After line 15 insert ": D = INV(D)"
	16	Alter "INV(D)" to "D"
Ch. 7 9p2	23	Alter "RSYMR1" to "RSYMRI"
	24	Move ":ERROR = SQRT(ERROR)" below comment ending "COEFFICIENTS."
Ch. 7 5p3	3	Alter "AAE" to "AGE"