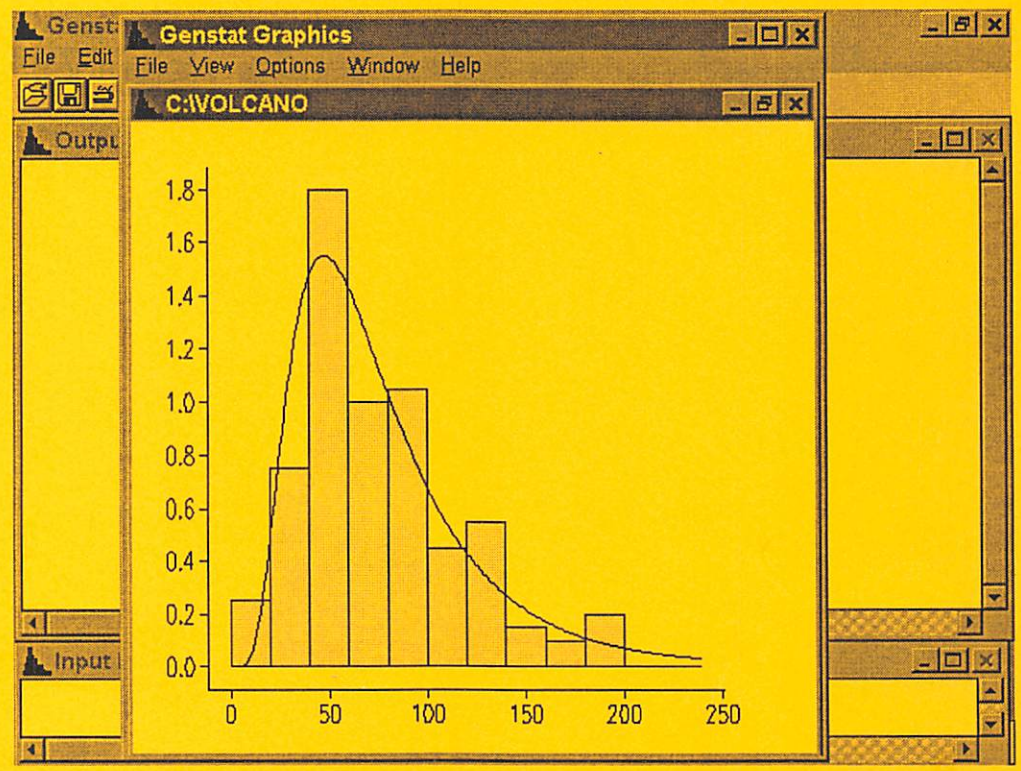
NAG®

GENSTAT

Newsletter

Issue 34



Editors

Sue Welham
AFRC Institute of Arable Crops Research
Rothamsted Experimental Station
HARPENDEN
Hertfordshire
United Kingdom AL5 2JQ

Anna Kane
NAG Ltd
Wilkinson House
Jordan Hill Road
OXFORD
United Kingdom OX2 8DR

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

ISSN 0269-0764

NAG Bulletin Board:

http//www.nag.co.uk/

Genstat Newsletter

Issue 34

NP3248

Jan 1998

Contents

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

Editorial

Welcome to the final release of the Genstat Newsletter in its present form. As mentioned in the Newsletter 33 Editorial, the Genstat Newsletter is evolving and henceforth users will receive an annual technical journal together with a more frequent and more modern, NAG statistical bulletin.

To round the successful series of Genstat Newsletters off, there is a generous helping of Genstat Talk and numerous articles illustrating a wide range of useful Genstat applications. As usual, the code for any procedure listed in any of the articles may be found on the NAG bulletin board.

The editors would like to take this opportunity to point out that the 3rd Edition of Genstat for Windows will soon be available and anyone wishing to view the enhancements to the system can do so on the internet at

http://www.nag.co.uk/stats/TT/demo/win3ed.html.

GENSTAT TALK

Extracts from the Genstat electronic discussion list, January to December 1995, summarized and edited by Peter Lane, Rothamsted. To join the discussion, send the message:

SUBSCRIBE Genstat *first-name last-name*

to the address: LISTSERV@LISTSERV.RL.AC.UK

Messages are archived monthly, and can be retrieved by sending, to the same address, a message like

SEND Genstat log9501

to get the archive for January 1995. You can also search the archives on the Internet (excluding the current month) by connecting a web browser to the URL:

gopher://jse.stat.ncsu.edu:70/11/othergroups/genstat

The opinions expressed here are not necessarily endorsed by either NAG or Rothamsted, and statements may not have been checked for accuracy. However, members of the Genstat development team and of NAG's Statistics Section are contributors to the discussions.

Unbalanced analysis

Query: I'm analysing data which are not completely balanced and I therefore use the FIT directive. Since my data are arranged in a split-plot design and hence different terms need to be tested against different residual mean squares, I usually calculate F-values by hand. Is there another way to handle this problem?

meyerl@ubaclu.unibas.ch

Reply: Wouldn't REML do what you want? An unbalanced multi-stratum design is easily handled by REML (look at VCOMPONENTS and REML in the Manual), or do you not want to combine information across different strata?

bairdd@agresearch.cri.nz

Follow-up: You can use the REML algorithm, which fits mixed models to unbalanced data using the residual maximum-likelihood method. In this case, you can use Wald statistics to assess the effects of each model term. For example, where for balanced data you would use the commands:

```
BLOCKS block/main/sub
TREATMENTS trt
ANOVA y
```

for an unbalanced data set you can use:

```
VCOMPONENTS [FIXED=trt] RANDOM=block/main/sub
REML [PRINT=model,comp,wald,mean] y
```

See Chapter 10 of the Release 3 Manual for details.

sue.welham@bbsrc.ac.uk

Generalized linear mixed models

Query: Has anyone tried to do something formal in model selection with GLMMs? Along with random effects, I'm fitting three fixed covariates (as many as five levels per covariate) and I would like to test whether specific interactions define variation. Any suggestions?

esa151@sac.ed.ac.uk

Reply 1: I wanted to test two fixed effects and their interaction in a Poisson/log-link GLMM recently. I noticed that unlike REML, the GLMM procedure does not give any tests of fixed effects. I loaded GLMM and edited it so that at the last iteration of Schall's iterated REML algorithm I got REML to print the Wald statistics. I'm not sure of the validity of this. Alternatively, the Welham & Thompson likelihood-ratio test for linear mixed models could be used, but that's a more tricky change to GLMM since you have to pass in the submodel being tested. Also I believe that the change-of-deviance test could be used by fixing the random effect variances when running the fixed-effect submodel. That doesn't appear to be implemented yet in REML. Again I don't know if these are valid in the GLMM context: I'll leave that to the experts.

sgc@forestry.tas.gov.au

Reply 2: Jeff Wood has suggested an easier way to get the Wald tests after a GLMM fit. You can use VDISPLAY [PRINT=wald]. Much simpler than my suggestion of editing GLMM.

sgc@forestry.tas.gov.au

Addendum: I think that Sue Welham deserves the credit for this. I learnt a lot from her visit to CSIRO.

jeff@canopy.biom.csiro.au

Cross-tabulation

Query: I am trying to put together a prototype program for our scientists to produce cross-tabulations of frequencies from survey data. The way I have been doing this is to use the TABULATE directive. It requires you, however, to tabulate on a variate and ask for the count of the number of units for that factor combination. Here is an example of what I would like to improve on:

```
VARIATE [VALUES=10(1)] dummy
TABULATE [CLASS=q1,q2; COUNTS=c12] dummy
```

franzp@agvic.gov.au

Reply: The Release 3 Manual says on Page 215 that "The data values are irrelevant for counts, and so you do not need to list any if counts are all that you require." So the dummy variate is not required:

```
TABULATE [CLASS=q1,q2; COUNTS=c12]
```

hendersonh@ruakara.cri.nz

Estimate of a ratio of parameters

Query: I am trying to use the procedure FIELLER to get a confidence interval for a ratio and I am supplying estimates and the variance-covariance matrix via the options. The only problem is that I cannot see how to pass across appropriate degrees of freedom for the t-deviate which is involved in applying Fieller's theorem.

jeff@canopy.biom.csiro.au

Reply: FIELLER is set up to deal specifically with the analysis of quantal data. The value of "t" it uses is therefore taken directly from the Normal distribution, inside the procedure. You could modify this by setting a new option to supply a value for t. Use

```
LIBEXAMPLE 'FIELLER'; SOURCE=s
OPEN 'fieller.gpi'; CHANNEL=2; FILE=out
PRINT [CHANNEL=2] s; JUSTIF=left; SKIP=0
```

to extract the procedure from the Library, and then edit the resulting file.

peter.lane@bbsrc.ac.uk

Case in commands

Query: i have tried to run the example on page 247 of the genstat manual, but it tells me that the values for method have not been set. any suggestions?

```
Genstat 5 Release 3.1 (Sun/Unix) ...
   PROCEDURE 'TRANSFORM'
   option name='method'; mode=t; values=!t(logit ...
   parameter name='percent','result'; ...
   if method.eqs.'logit'
         calc result = ...
   elsif ...
   ...
   ENDPROCEDURE
   TRANSFORM [method=a] every10%; result=logit10%
******** Fault (Code VA 4). Statement 1 in Procedure
TRANSFOR
Command: if method.eqs.'logit'
Values not set
method has no values
```

simon@bioss.sari.ac.uk

Reply: You have run into a problem with case. I would have thought that someone evidently using a Unix machine would be as case-sensitive as Genstat! The rule is that the dummy identifiers you use to refer to options and parameters within the body of a procedure must be in upper case: see Page 245. The example you quote on Page 248 does actually refer to "METHOD" rather than "method".

peter.lane@bbsrc.ac.uk

(Also answered by *potter.bbsrc.ac.uk* and *simon.harding@bbsrc.ac.uk.*)

Follow-up: It is a great shame that Genstat is so old-fashioned and inconsistent in this respect. The rest of the program is case-insensitive, and newcomers to it (whether they are Unix users or not) are right to expect case not to matter.

tim.cole@mrc-dunn.cam.ac.uk

Rejoinder: I don't think that this is fair comment. What is old-fashioned about being case-sensitive? It is the old-fashioned software, and hardware, that ignored case; even now, the Fortran 77 standard insists on all statements being in upper case. Unix seems to be regarded by many people as the operating system of the future, and it is certainly case-sensitive.

peter.lane@bbsrc.ac.uk

Response: Thank you for the reply regarding the procedure example; it was indeed related to the use of upper casing, something which i am unused to, having used pc's until now. I shall pay a bit more attention to such details in the future!

simon@bioss.sari.ac.uk

Missing values in time series

Query: One of my colleagues has a time series with about 600 values, of which about 30 are missing. He has estimates of the missing values using TKEEP. What is the most elegant way to substitute them into the original series? We did it using RESTRICT to find the locations of the missing values, but this seems clumsy.

jeff@canopy.biom.csiro.au

Reply 1: Can't you use

```
CALC X = MVREPLACE(X; Y)
```

for this? See Page 149 of the Manual. You'll have to fill up Y with some stuffing, which makes the solution less elegant. The source code for the MULTMISS procedure may give some neat ideas.

sedcole@whio.lincoln.ac.nz

Rejoinder: It seems to me that this is more complicated than the original use of RESTRICT to get a variate, Mvno say, of positions of the missing values and then using

```
CALC X$[#Mvno] = Y
```

jeff@canopy.biom.csiro.au

Reply 2: There is an option MVREPLACE of ESTIMATE which allows replacement of missing values in a time series by their estimates after fitting a model. You would usually use this option when, after trying several models, you had selected the best one. Of course, ESTIMATE does the work of locating these missing values, so it would be possible for TKEEP to supply the position vector for them. If a transfer-function model is fitted, the missing values and their s.e.s are presently returned in one vector for the input and output series together, but it would be possible for TKEEP to put these and the positions into pointer structures.

tunnicliffe-wilson@lancaster.ac.uk

Generalized estimating equations

Query: I am new to this list. I wonder if there is any program for Generalized Estimating Equations (GEE) applicable with Genstat. If so, how can I get it?

akhtar-danesh@newcastle.ac.uk

Reply: Yes, there is, because I wrote procedure GEE in conjunction with Mike Kenward. It will be appearing in Procedure Library 3[3] with Release 3.2 of Genstat, and a couple of articles have appeared recently in the Genstat Newsletter.

dsmith@neumann.une.edu.au

Printing expressions

Query: The following code illustrates that a single value text and an expression cannot be printed in parallel without first creating a textual version of the expression. Although only a minor point, with a work around, I wonder if anyone has a way of printing them in parallel directly?

```
TEXT T; 'Expression used :'
EXPRESSION E; !e( 2 * N - 4 )
PRINT T,E
```

This prints them in series. A work-around is to create a text version of the expression:

```
PRINT [CHANNEL=Text_E; IPRINT=*; SQUASH=yes]
E
PRINT T, Text_E
```

peter_coleman@sandwich.pfizer.com

Reply 1: The following gives a parallel print:

```
PRINT [IPRINT=*] T,E; SKIP=*
```

anna@nag.co.uk

Follow-up: The ability to print expressions (and formulae for that matter) is not covered fully in the Manual. I looked into this earlier in the year in response to a direct request, and attach the gist of what I discovered [extract the archive for June 1995 to get this report]. In summary, expressions can only be printed in series; but, as noted in the previous response, the SKIP parameter has the (unexpected) effect of suppressing newlines in serial printing with formulae and expressions as long as the option IPRINT=* is set.

peter.lane@bbsrc.ac.uk

Confidence limits for predictions

Query: I want the confidence limits for an individual prediction $\hat{y}$ at an explanatory value x. The PREDICT directive will give me the s.e. to work out the limits for mean($\hat{y}$) at x. How do I get Genstat to reflect the error as well as the variance in the parameter estimates, apart from extracting the results of the fit myself and using CALCULATE?

callinanl@goldy.agvic.gov.au

Reply: It may be simpler to use the work already done by PREDICT, as follows:

```
MODEL y
FIT x
PREDICT [PRED=pred; SE=sepred] x
RKEEP DEV=dev; DF=df
CALC seind = SQRT(sepred**2+dev/df)
```

h.van.der.voet@glw.agro.nl

Editor's note: During the time since this question was asked, the PREDICT directive has been extended to solve this problem generally. In Release 4.1, there is an option, SCOPE=new, which allows you to add in the contribution from the error variance in any linear model or GLM.

Rounding in calculations

Query: I have a very simple question and have only been able to answer it using complicated contortions. I have five plots of eight trees each. Tree heights per tree are recorded in a single column made up by Trees 1-8 of Plot 1, then Trees 1-8 of Plot 2, and so on. Now I would like to re-arrange the data so that it comprises eight variates on each of five plots. A logical idea which does not work is:

```
UNITS [40]
OPEN 'tree.dat'; CHANNEL=2
READ [CHANNEL=2] X
UNITS [5]
EQUATE OLD=X; NEW=Y[1...8]
```

clarkep@stat.unp.ac.za

Reply 1: This does not work because the parameter for NEW must already have been declared as a pointer. I am not clear whether you want five variates of length 8 or eight of length 5. For the first, use

```
VARIATE [8] Y[1...5]
EQUATE OLD=X; NEW=Y
```

and Y[*n*] will contain the eight trees in Plot *n*. For the second, use

```
VARIATE [5] Y[1...8]
EQUATE [OLDFORM=!(((1,-7)5,-1)8)] OLD=X; NEW=Y
```

and Y[*n*] will contain the values for Tree *n* in each of the five plots.

ian@sass.sari.ac.uk

Reply 2: There is an even simpler solution using READ, assuming that the data do not need to be stored in a single structure as well as in the separate structures. To get one variate for each plot, use

```
OPEN 'tree.dat'; CHANNEL=2
VARIATE [8] Y[1...5]
READ [CHANNEL=2; SERIAL=yes; END=*] Y[]
```

Alternatively, to get one variate containing measurements on the first trees in each plot, one for the second, and so on:

```
OPEN 'tree.dat'; CHANNEL=2
READ [CHANNEL=2] Y[1...8]
```

peter.lane@bbsrc.ac.uk

Percentiles

Query: I want to find the 97.5 percentile point of a series of numbers. Is this possible with the only genuine Statistics package, and how?

kerym@ubaclu.unibas.ch

Reply 1: Procedure QUANTILE does this:

```
QUANTILE [PROPORTION=0.975] variate
```

peter.lane@bbsrc.ac.uk

Further discussion: There were some queries about different definitions of percentiles, and their availability in other packages (July 1995).

GLMM and HGLM

Query: I have used GLMM with a log link (with Poisson and sometimes gamma) and have the impression that the back-transformed means still have an appreciable bias. I cannot discern from the code what is being done, but wonder if more than a first-order correction for the residual mean square may be of benefit.. This suggestion arises from the remark in the last paragraph of Section 6 of the Library Manual: "Initial values for the variance components are calculated by REML estimation using the fixed and random models ON THE DATA TRANSFORMED BY THE LINK FUNCTION".

llefkovi@ccs.carleton.ca

Reply 1: Have you tried the option in GLMM for fitting the marginal model of Breslow and Clayton (FMETHOD=fixed)? The default is the subject-specific model fitted using Schall's algorithm. The SS fixed-effect estimates need to be adjusted if you want to predict the marginal mean (see Zeger et al 1988 Biometrics 44 1049-1060); alternatively you could fit the marginal model directly.

sgc@forestry.tas.gov.au

Reply 2: The solution is to use the Poisson-gamma HGLM in place of the Poisson-Normal GLMM. With the former you have mean(fitted values) = mean(data) so that there is no bias. The HGLM assumes that Y is Poisson with mean mu'=mu*u where u has some gamma distribution with mean 1 and shape parameter nu. If anyone wishes to try out these models, I have Genstat code for one, two or three random components with the conjugate models Poisson-gamma, binomial-beta, gamma-inverse gamma, and Normal-Normal plus the GLMMs Poisson-Normal, binomial-Normal and gamma-Normal. Each set of procedures has a manual, and all require the K- system to be loaded first. They are not guaranteed to be bug-free, so that offers to try them out would be appreciated! I can send email versions to anyone interested. Please say if you want the K- system as well.

j.nelder@thor.ma.ic.ac.uk

Multiple comparisons

Query: A researcher here wants to analyse the results of a survey to see if there are significant differences between the answers given by different demographic groups. Because there are a ton of questions to be analysed, we're looking for an approach which will not only give significance levels for the factors, but also tell us which levels are significantly different. Unfortunately, we've found that sometimes an effect (often an interaction) has a significant F-ratio, but the multiple comparison tests don't identify any significant differences between means. Can anyone please suggest a technical approach that will give consistent F-tests and multiple comparisons, an explanation for the inconsistency which I can give to the researcher, or best of all, something I can read to get a better idea of what is going on.

duncan.hedderley@bbsrc.ac.uk

Reply 1: Well done! By mentioning the phrase "multiple comparisons" you seem to have silenced everyone on the Genstat discussion list! This subject is a can of worms which is not lightly opened. Nevertheless, the "inconsistency" of multiple comparison procedures (MCPs) is a favourite hobbyhorse of mine, so I'll take the bait. I refer you to my 1990 paper in the American Statistician, 44, 174-80, and the subsequent replies in 1991, 45, 165-167. In a nutshell, my proposal is to ignore the overall F-test completely and make each pairwise comparison using an LSD for the comparison which is obtained by multiplying the SED for the comparison by the t-value. I should hasten to add that more general contrasts should be used if appropriate.

savilled@agresearch.cri.nz

Reply 2: My objection to this is that it is a general prescription for multiple situations. Really the issue ought to be: "What is the appropriate yardstick against which to measure the error rate: by experiment, by individual comparison, or by something else?". Any inconsistency resides in the MCPs, not in the tests themselves. It is entirely legitimate to test whether the max and min results are significantly different in an ANOVA. This is an MCP (Tukey, alias the omega test). Often it is appropriate to quote both the individual comparison LSD and the Tukey LSD. This allows the reader to weigh the evidence from both points of view. The LSDs are no more than yardsticks, to be used as we find them useful. Problems arise when we try to assign some higher status to them.

john@maths.marc.cri.nz

Editor's Note: this issue was thoroughly revisited later, with many statisticians contributing their views from round the world. To follow the full argument, get the archives for June and July 1995, or search for "Multiple Range Tests".

Saving the RSS from ANOVA

Query: My problem is how to save the residual sum of squares from ANOVA. The Manual says I should use

```
AKEEP "Units"; SS-name
```

but I cannot understand this well, so would appreciate it if somebody could explain where and how I could define "Units" and maybe give me another solution.

smoljanovic@maths.hull.ac.uk

Reply: The name of the units term in an ANOVA is set up in the UNITS statement; for example,

```
UNITS plot
```

In the absence of a UNITS statement declaring the name of the factor which indexes the units, Genstat allows you to reference the unnamed unit structure with the special name "Units" which is what you have done. If you name your units factor as above, then you can use a more explicit call to AKEEP:

```
AKEEP Plot; SS-ss
```

baird@bbsrc.ac.uk

Partial correlations

Query: I have a normal multiple regression model and want to get the partial correlation coefficients between the response variable Y and the regressors X(i). This is not possible with the CORRELATE directive. Does anybody know the simplest way to get them?

kerym@ubaclu.unibas.ch

Reply: I have written a program to do partial correlations, I think in Genstat but it may be Fortran. I'll dig it out and send it. I hope to turn it into a procedure one day.

potter@bbsrc.ac.uk

Editor's Note: The code, with an example, was later sent to the list, so can be retrieved from the archive, by sending the command

```
send genstat log9504
```

to the adddress

listserv@listserv.rl.ac.uk

Quantal assay

Query: I have two questions, both involving quantal assay. (1) When I use procedure FIELLER, I sometimes would like a series of confidence limits calculated, not just one. I have tried

```
FIELLER [LINK=probit] %DOSE=!(40,50,60,90)
```

and get an error message about incompatible numbers of values. (2) I am analysing the relationship between number of undamaged blood cells and dose of an anti-malarial drug. The values of the response typically start at about 5000 and follow a decreasing sigmoidal curve to nearly zero. Fitting a GLM with DIST=binomial; LINK=probit, I need to specify the number of binomial trials, which I usually take as the first observed response. Obviously this is very approximate: how can I improve on this?

clarkep@stat.unp.ac.za

Reply 1: Fitting sigmoid curves with an unknown upper (or lower) asymptote does not fit in the framework of GLMs. These models are truly nonlinear, and therefore the FITNONLINEAR directive should be used. Happily, Genstat allows us to use non-Normal error models (e.g. Poisson) also in this context.

h.van.der.voet@glw.agro.nl

Reply 2: I think the procedure PROBITANALYSIS will solve both your problems.

ccsphc@bath.ac.uk

Reply 3: I think that Question 2 is Wadley's problem, the analysis of which is available in Genstat procedure WADLEY.

potter@bbsrc.ac.uk

Reply 4: In response to Question 1, the procedure FIELLER needs a redesign in the light of the new features for procedures in Release 3. In the meantime, you can do what you want by setting the first parameter as well as the %DOSE parameter:

```
FIELLER TREAT=!(4(1)); %DOSE=!(40,50,60,90)
```

peter.lane@bbsrc.ac.uk

Rejoinder: Many thanks to the respondents to my queries. My limited experience has been that PROBITANALYSIS converges in several situations when WADLEY diverges. However, the greater opportunities such as in modelling overdispersion of WADLEY make it attractive.

I would like to see confidence (or fiducial) limits calculated for the LD50s in this situation, but of course the third parameter makes it hard. A colleague and I are looking at this problem.

clarkep@stat.unp.ac.za

Customized syntax of commands

Query: I've been going through an interactive model-fitting exercise and I am enjoying using SETOPTION to save a bit of typing. I thought I'd try SETPARAMETERing my TERMS, but the required formula doesn't have a name. Since I have a sequence of different MODEL statements but always the same terms, this is something I would like to be able to do. Is there any way around this?

littlejohnr@agresearch.cri.nz

Reply: Yes, it is possible to set a default even for the unnamed parameter of a directive with only one parameter. The Manual doesn't say how, but all you have to do is give a missing value in the NAME parameter of SETPARAMETER:

```
SETPARAM [DIR=TERMS] *; DEFAULT=!f(x1+x2*g)
```

There are very few directives with defaults for the first parameter: I think only ADISPLAY, RETURN, RKEEP and TKEEP. There must also be a few procedures, such as DAPLOT.

peter.lane@bbsrc.ac.uk

Formulae

Query: Why does the following not work

```
GLMM [RANDOM=rterm; FIXED=fterm] y; NBINOM=n
```

whereas it does work if I put # before rterm and fterm? I learnt long ago that if a Genstat job does not work, a sprinkling of #s will often do the trick. Sometimes I even understand why, but this particular example puzzles me. It reminds me of the old joke about Wirth (inventor of Pascal): "I do not care whether you call me by name or call me by value".

jeff@canopy.biom.csiro.au

Reply: The use of neither formulae nor expressions are spelt out in full detail in the Manual. The main problem is that the words "expression" and "formula" are used both to describe the pieces of syntax constructed with operators (Page 24) and the data structures used to store the constructions (Page 40). The word "formula" (as in the description of the GLMM procedure) generally means the construction, whereas "formula structure" means the structure identifier. The Genstat language allows the alternative of reference by value or by name only for scalars and quoted strings. I suspect that even Pascal does not provide this alternative for expressions.

peter.lane@bbsrc.ac.uk

Procrustes analysis

Query: Not long ago, we started using Procrustes analysis for the analysis of data generated from sensory evaluation experiments. At first, we were thrilled about the method. Then we came across a paper by Alan Huitson entitled "Problems with Procrustes analysis", 1989, J Appl Stat 16, no 1, in which he states that the method should be used only circumspectly. We also tried an analysis with simulated data, and found that results we got were almost as good as with our real data, though no treatment effect had been simulated! Has anybody ever experienced this?
n_rodrigue@qcrssh.agr.ca

Reply: Please note that a letter in reply to the article by Huitson was published in the Journal of Applied Statistics (1990, 17, 449-451). The so-called "problems" are essentially in misunderstanding the underlying assumptions of the method and are too superficial a presentation of results by the original author. This is not in any way to do with the Genstat 4 macro used by Huitson, nor the (improved, more extensive) Genstat *5* procedure GENPROC currently available in the Procedure Library.
gillian.arnold@bbsrc.ac.uk

Parallel predictions

Query: For a quadratic polynomial, I would like to obtain predictions and s.e.s for several new values of X. I would like to give PREDICT a parallel list of values of X and X^2, but as I understand it, PREDICT only gives predictions for every combination of X and X^2. I need only the diagonal elements. Does anyone have a way of fooling PREDICT into giving parallel predictions?
franzp@agvic.gov.au

Reply: I don't think there is a way of getting PREDICT to treat sets of values of explanatory values in parallel. However, the problem as posed is easily solved using the POL function rather than constructing your own polynomials. Instead of

```
CALC X2 = X**2
FIT X+X2
PREDICT X,X2; LEVELS=!(1,2,3),!(1,4,9)
```

you can (since Release 3.1) use

```
FIT POL(X; 2)
PREDICT X; LEVELS=!(1,2,3)
```

The PREDICT directive then has no problem with producing the required results. (If you use REG rather than POL to get orthogonal polynomials, I'm afraid that PREDICT is then unable to produce predictions at all.)
peter.lane@bbsrc.ac.uk

Landscape graphics

Query: Silly question, I am sure, but how do you get landscape graphics via a PostScript file out of Genstat 3.1? I can't see it anywhere, but I remember doing it some years ago in Release 2.
logsdon@lancaster.ac.uk

Reply: This command should do the trick:

```
DEVICE 4; ORIENT=land
```

It is still square, so does not fill up the page. Does anyone know how to do this?
p.baker@prospect.anprod.csiro.au

Reply to new question: You can do this by setting limits with the AXES directive:

```
DEVICE 4; ORIENT=land
AXES 1; XUPPER=1.4
```

and similarly for portrait. If you use landscape, the PostScript file may need editing (it does with my version of Genstat, anyhow). Near the top of the file, change

```
90 rotate -720 0 traslate} def
```

to

```
90 rotate 0 -540 traslate} def
```

otherwise you get a blank page when the file is printed!
butlerr@crop.cri.nz

Editor's Note: This topic continued with various graphical requests, and suggestions that tips like this should be more readily available. A repository was set up soon after at Statlib, and you can connect to it by the Internet at the URL:

http://lib.stat.cmu.edu/genstat

Extracting a subset of values

Query: I am trying to restrict a set of variates, concatenate them and use the new variate in a further analysis. I thought that I might be able to do this using the ELEMENTS function:

```
CALC z[1...10] = ELEM(x[1...10]; r<=0.5)
```

This does not work, giving a diagnostic because the expression returns the value 0, which is not a valid address of the structure. Is there a way that ELEMENTS can be used here? The alternative approach is to use RESTRICT:

```
RESTRICT y; r<=0.5; SAVE=sv
CALC z[1...10] = x[1...10]$[sv]
EQUATE OLD=z; NEW=w
```

bobf@candid.biom.csiro.au

Reply: I would just use the SUBSET procedure which I find extremely useful, especially for factors:

```
SUBSET [r<=0.5] OLD=x[1...10]; NEW=z[1...10]
```

p.baker@prospect.anprod.csiro.au

Kronecker products

Query: I am having difficulty in one step of trying to form a Kronecker product of two (rectangular) matrices. The easy part is to obtain each submatrix; the part I have failed to get to work is assembling the submatrices into their correct positions. I feel sure that there is an easy way, but it escapes me. As an example, I have a 6x6 non-symmetric matrix A whose Kronecker product with a 4x4 non-symmetric matrix B is wanted. In my current code, I multiply B by each element of A in turn, storing the results in 36 matrices, each 4x4. It would save space and time if each of these could be placed into their correct position in the 24x24 array.
llefkovi@ccs.carleton.ca

Discussion: There were several contributions to the discussion, which wound up with some code from *peter.lane@bbsrc.ac.uk* being encapsulated into a short and general procedure by *nelder@imperial.ac.uk*:

```
PROCEDURE 'KRONECKER'
PARAMETER 'A','B','C'; SET=y,y,n; PRESENT=y,y,n
CALC ra,rb = NROW(A,B) & ca,cb = NCOL(A,B)
& rc,cc = ra,ca*rb,cb
MATRIX [rc; cc] C
VARIATE [rb] ri[1...ra] & [cb] ci[1...ca]
EQUATE OLD=!(1...rc),!(1...cc); NEW=ri,ci
CALC C$[ri[#ra(1...ra)]; ci[(1...ca)#ca]] = \
    #A$[#ra(1...ra); (1...ca)#ca]*B
ENDPROCEDURE
```

Ignoring fatal errors

Query: I am trying to bootstrap a function of parameters from an inverse linear model. Unfortunately, sometimes, within a loop which does the resampling, the analysis does not converge and a fatal error terminates the process. Can I arrange to ignore fatal errors and continue?
jconnoll@statlan.ucd.ie

Reply: You can continue execution after a fault by

```
SET [DIAGNOSTIC=*]
```

This suppresses the error message and the subsequent abandonment of the job in batch mode. You can see the error message with

```
DISPLAY
```

but remember that you will not see any reports of "messages" or of "warnings" either, and you cannot recover those from DISPLAY.
peter.lane@bbsrc.ac.uk

Page scrolling

Query: Does anyone know how to switch off page scrolling in Release 3.1, so that the output comes out in one chunk? I've tried setting PageScroll to 0 in the Genstat configuration file (.g31config on a Unix system), but that seems to give the default of 16 lines of output before prompting for a Return. One approach, I assume, is to set PageScroll to a large number. This still risks the appearance of a prompt when the output is extensive. We use the Genstat-Emacs interface, and prefer to scroll back and forth at will through the whole of the output from the latest set of directives.
john@maths.marc.cri.nz

Reply: The statement

```
SET [PAUSE=0]
```

in the start-up file (startup.gen in the support directory) will have the required effect.
anna@nag.co.uk

Follow-up: The PageScroll setting is one of the controls of the Interacter interface for Genstat (common to the PC, Vax, Sun and Alpha versions). It controls the number of lines treated as a "page" when you press the PageUp or PageDown key to look back at previous output. It can be set either by entering Setup Mode interactively (by pressing F3) or by editing the configuration file. When not reviewing output, the number of lines sent to the screen before an automatic pause is controlled by the PAUSE keyword of SET, as explained in the first reply.
peter.lane@bbsrc.ac.uk

Designing experiments

Query: I find the DESIGN procedure in Genstat to be an extremely useful method of generating experimental designs. However, since the design procedure can only be used interactively, how does one obtain a hard copy of the design?
hohlst@gene.unp.ac.za

Reply 1: The blocking and treatment structures (and unit factor) are named in the procedure DESIGN. You can open an output file and print the structures. Also, in the procedure PDESIGN, you may name the table and print it to the output file.
ken@sass.sari.ac.uk

Editor's note: There was more discussion of the use of the DESIGN procedure, which can be found in the archive for August 1995.

Duplicate structures

Query: I am having difficulty understanding what the DUPLICATE directive does. The manual talks about "the attributes to be duplicated", but when I overwrite the old structure and its attributes, I get an error message which suggests that the attributes of the new structure are being overwritten as well. This suggests to me that the attributes were not duplicated after all.
jeff.wood@cbr.biom.csiro.au

Reply: The intention is to form a structure which is either a copy of an existing one (the default), or which has some selected attributes the same. The reported behaviour results from the fact that the attributes themselves are not duplicated, only the structures themselves. Take a simple example, duplicating a factor with a set of labels:

```
TEXT [VALUES=a,b] t
FACTOR [LABELS=t] f
DUPLICATE OLD=f; NEW=g
```

This sets up g to be another factor with labels defined by the text structure t. If you now delete t, Genstat warns you that both f and g are affected because they both rely on t.
peter.lane@bbsrc.ac.uk

Suppressing new pages

Query: I give in! The statement

```
SET [OUTPUT=*]
```

suppresses the new page and dots associated with a FIT directive. However,

```
SETOPTION [SET] OUTPUT; *
```

in the start-up file seems to have no effect. No other alternative seems to give the desired effect. What is the logic of this?
lschmitt@anhb.uwa.edu.au

Reply: You are trying too hard! If you want to suppress the new page and dots associated with FIT and other directives, just put the statement

```
SET [OUTPRINT=*]
```

in the start-up file. The statement

```
SETOPTION [SET] OUTPUT; *
```

has the effect of defining the default for the OUTPUT option of the SET directive: it has no effect directly on the output that is produced by other directives. You have to use the SET directive subsequently for the action to occur. So you could follow up with

```
SET
```

to get the required action, either in the start-up file or in the course of a program.
peter.lane@bbsrc.ac.uk

Workspace needed for REML

Query: I am using REML with a view to calculating genetic correlations, and keep getting a message that there is insufficient space available to form Wald statistics or to check the variance matrix. This sounds like it might be serious. Is it?
jeff@canopy.biom.csiro.au

Reply 1: I've had similar problems when analysing Alpha designs. I think that things should be better with the new version, particularly on PC.
potter@bbsrc.ac.uk

Reply 2: Wald statistics are formed from a Cholesky decomposition of the $X'V^{-1}X$ matrix, which has the side effect of checking that the variance matrix is positive definite (which is only checked implicitly elsewhere, since in order to save space the V matrix is never formed in full during iterations). If all the components are positive, there is no problem. If some are negative, then usually there is no problem, but as positive-definiteness for new estimates can only be checked on the next iteration, there is a small chance that estimates will go out of bounds on the final iteration. Since changes are small at convergence this is unlikely, but it does occasionally happen, hence the warning. This is explained on Page 571 of the Manual.
sue.welham@bbsrc.ac.uk

Editor's note: More discussion followed on the implementation of REML in different releases and on different operating systems (July 1995).

Some procedures for analysing repeated measures in Genstat

Roger Littlejohn
AgResearch, Invermay Agricultural Centre
MOSGIEL, New Zealand

littlejohnr@agresearch.cri.nz

1. Introduction

This note describes two Genstat procedures I have developed for analysing repeated measures data. Firstly, RMSUMM (Repeated Measures SUMMary) for obtaining summary tables and graphs, and secondly, RMBAL (Repeated Measures for BALanced data sets) for fitting the model described in Chapter 5 of Diggle (1990) when measurements from each unit are available at each time. I will use Box's rat weight data (Box, 1950) for illustrative purposes.

The repeated measures data I get falls into several categories, each requiring its own approach.

1 Growth curves, as in live weight measured over days, weeks, months or years.
2 Nonlinear response curves driven by differential equations.
3 Profiles for which some *a priori* theory suggests a suitable time contrast or model.
4 Profiles for which "no specific features are known to be of interest *a priori*" (Kenward, 1987), but a comparison between treatments is required.

Some data sets are multivariate: for example, parallel sets of glucose and insulin profiles. The example used falls in the first category.

Given that "repeated measures" covers a broad range of types of data and there is no one right method of analysis, what are the options? I am familiar with the following approaches to analysing repeated measures, which are available in Genstat.

1 Univariate ANOVA at each time, for estimation purposes, rather than testing.
2 Univariate ANOVA of summary statistics, including means, maxima or minima, linear contrasts, orthogonal polynomials (VORTHPOL), estimated curve parameters (derived perhaps from FITCURVE, FITNONLINEAR or FITSCHNUTE) and area under the curve.
3 Analysis of ante-dependence, with subsequent ANCOVA (ANTORDER, ANTTEST).
4 Split-plot ANOVA with a flat covariance matrix, and polynomial or nonlinear contrasts in time (NLCONTRASTS) or possible modification of F degrees of freedom (REPMEAS).
5 Use of a modelled covariance matrix, with REML or RMBAL (ML).
6 MANOVA, with a full covariance matrix.
7 Generalized estimating equations.

2. Graphs

Graphs are the starting point for most statistical analyses, particularly so for repeated measures. The most important graphs are (i) individual profiles grouped and apart, to check for homogeneity of variance and outlying values, (ii) treatment group means over time to assess trends and patterns, and (iii) some expression of correlation as a function of time lag.

RMSUMM plots (Figure 1) and tabulates the means for each level of a factor with the SED for each time:

```
BLOCK
TREAT treat
RMSUMM [PR=unadjusted; TERM=treat] Y=y[]; TIMES=0...4
```

Times	control	thyroxin	thiouracil	SED	V.R.	Fprob
0	3.984	4.016	3.998	0.04240	0.29	0.754
1	4.356	4.326	4.330	0.05310	0.22	0.805
2	4.659	4.648	4.559	0.04752	3.16	0.061
3	4.864	4.881	4.682	0.05210	10.11	0.001
4	5.075	5.085	4.818	0.05204	19.28	0.000

```
RMSUMM [PR=adjusted; ORDER=2; TERM=treat] Y=y[]; TIMES=0...4
```

Times	control	thyroxin	thiouracil	SED	V.R.	Fprob
0	3.984	4.016	3.998	0.04240	0.29	0.755
1	4.371	4.306	4.329	0.02832	2.77	0.085
2	4.642	4.661	4.566	0.02690	7.65	0.003
3	4.811	4.827	4.773	0.02168	2.03	0.156
4	5.014	5.002	4.937	0.02552	3.07	0.067

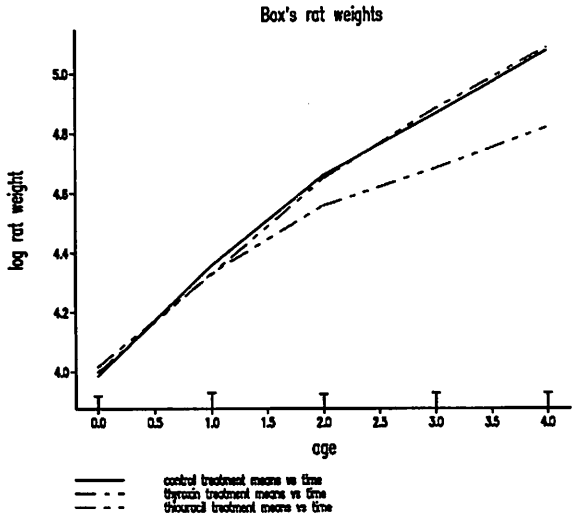


Figure 1(a): Mean log weight over age for each treatment group for Box's (1950) rat data, with covariate adjustment for previous 2 time points.

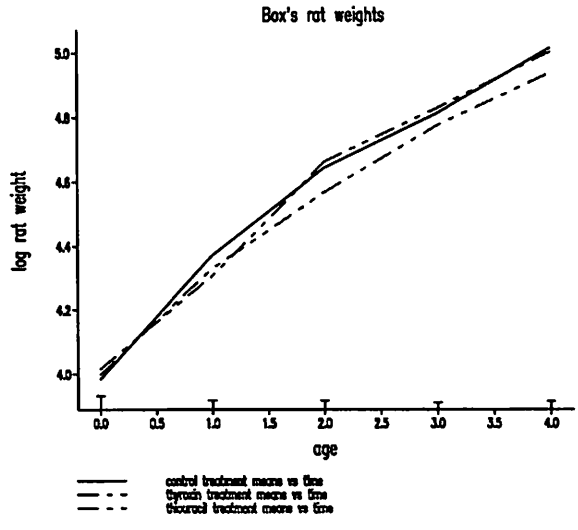


Figure 1(b): Mean log weight over age for each treatment group for Box's (1950) rat data, with SEDs for each age.

The (max-min) SEDs are given for unequal replication. RMSUMM differs from DREPMEAS in that (i) it uses ANOVA rather than TABULATE to obtain the means and thus has access to the SEDs, and (ii) it currently accepts only one factor rather than two. RMSUMM may do a covariate adjustment of specified order (suggested perhaps by ANTORDER), and provide the corresponding table and graph. In the course of time I may integrate REML into the procedure alongside ANOVA.

Further graphs are provided by RMBAL, in particular, a graph of the variogram and separate graphs for each level of a factor of the profiles for each unit (Figures 2 and 3). The variogram does not depend on equal spacings between samples (as in time series analysis), and is basically the variance (at lag 0) × (1 - the correlogram). RMBAL calculates a pooled variogram. The main points in interpreting it here (Figure 2) are that the gap at the bottom corresponds to the measurement variance, the gap at the top corresponds to experimental unit variance, and the shape of the variogram may suggest a way of modelling it.

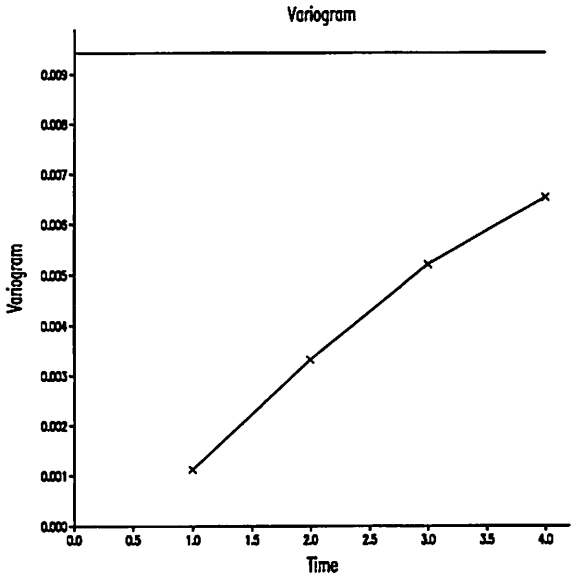


Figure 2: Variogram for log weight as output by procedure RMBAL.

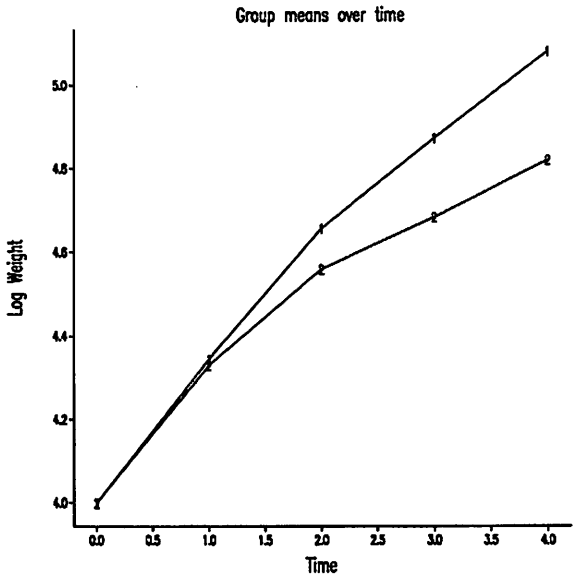


Figure 3(a): Mean log weight for treatment groups (combining control and thyroxin treatments).

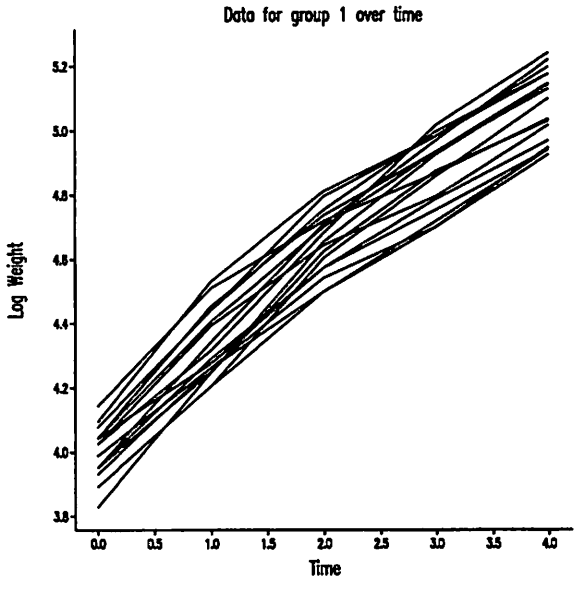


Figure 3(b): Individual log weights for control and thyroxin groups.

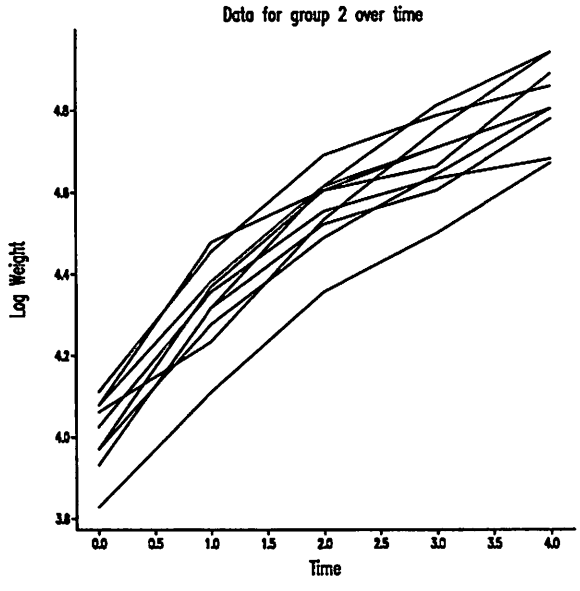


Figure 3(c): Individual log weights for thiouracil groups.

3. RMBAL

The model described in Chapter 5 of Diggle (1990) is constructed as follows. The response pattern over time is modelled by a polynomial trend or covariate, with possibly different parameters with treatment. The empirical variogram is used to suggest a suitable correlation structure within each time sequence, which is then modelled with parameters for measurement error, variation between experimental units, and serial correlation within units.

The model is fitted by maximizing the likelihood, written as a Genstat EXPRESSION, using FITNONLINEAR.

This model may in fact be specified in Genstat by setting the COVARIANCEMATRIX parameter of the VCOMP directive to the corresponding covariance matrix, and using REML. The equation in Section 10.2.5 of the Genstat 5 Release 3 manual

$$V = \sigma^2(\gamma_b Z_b I_3 Z_b' + \gamma_p C + I_{18})$$

is parametrized differently from Diggle's in Chapter 5.5, with

Diggle	Genstat
$\tau^2 = \phi_1 \sigma_d^2$	σ^2
σ_d^2	$\gamma_p\sigma^2$
$\nu^2 = \phi_2 \sigma_d^2$	$\gamma_b\sigma^2$

so that $\gamma_b=\phi_2/\phi_1$, $\gamma_p=1/\phi_1$. Also, Genstat would tend to use $\beta=e^{-\alpha}$ rather than α for correlation. When REML takes parametrized covariance matrices, RMBAL will be redundant.

The specification of the options and parameters is as follows.

Options

PRINT = *string*	What to print (summary, estimates, correlations, monitoring); if *, PRINT is suppressed; default estimates.
GROUPS = *factor*	Defines grouping of the units.
VARIOGRAM = *strings*	When to output variogram in relation to model fitting (before, after); if *, VARIOGRAM not output; default before.
CORRELATIONSTRUCTURE = *string*	Specifies correlation structure for variance matrix (exponential, gaussian); default gaussian.
MODELSTRUCTURE = *string*	Specifies nested models for likelihood analysis (full, intercept, parallel, null); default full.
TREND = *value*	Polynomial order of trend (0, 1, 2, 3, 4, 5); default 1.
COVARIATES = *pointer*	Contains covariates at TIMES.
INITIAL = *variate*	Contains initial parameter estimates as !(phi2,phi1,alpha).
CONSTRAINTS = *string*	Whether to set between- or within-unit components of variance to zero (between, within); default *.
GRAPHICS = *string*	Style of graphical output (lineprinter, highquality);
YTITLE = *text*	Title for *y*-axis of graphs; default "Response".
XTITLE = *text*	Title for *x*-axis of graphs; default "Time".
SAVE = *pointer*	Saves parameter estimates.

Parameters

DATA = *variates*	Data variates (observed at successive times).
TIMES = *scalars*	Times at which the variates are sampled.

A linear correlation structure is available in an updated version of RMBAL, which is available on the NAG Web page.

Output for the variogram and model fitting analysis of the log weights for Box's rat data are given below. (Note that the factor digtreat combines levels control and thyroxin of treat.) The data variates are given in DATA and the times at which data are sampled should be given in TIMES; currently this must be set, so that any set of DATA suffixes can be accessed. The data should be classified by one factor, given in GROUPS, and higher levels of blocking structure will be ignored. The data should not be restricted or contain missing values. Graphs

of the group means over time, and of the data over time for each treatment level, are controlled by the GRAPHICS, YTITLE and XTITLE options. These can be used to assess the need for a transformation and to obtain a suitable form of response over time, which is then specified in TREND as a polynomial of degree no greater than four. COVARIATES may also be included in the analysis; for the example given below, the quadratic TREND has been expressed through the COVARIATE option, facilitating more flexible modelling.

The VARIOGRAM option allows the variogram to be graphed either before model fitting, or after model fitting together with the fitted variogram based on the full model. This aids the choice of exponential or Gaussian functions for correlation over time, specified with the CORRELATIONSTRUCTURE option, and of initial values of phi2, phi1 and alpha, which may be given in the variate INITIAL (overwriting rather loose values calculated internally).

The GROUPS factor may be interpreted in six ways, controlled by the option MODELSTRUCTURE:

1.	MODELSTRUCTURE = *full*	The polynomial trend plus covariate model is fitted for each level of the factor.
2.	MODELSTRUCTURE = *intercept*	A common intercept is fitted for each level of the factor, but other parameters are separate.
3.	MODELSTRUCTURE = *covariate*	A common covariate is fitted for each level of the factor, but other parameters are separate.
4.	MODELSTRUCTURE = *covint*	Common intercepts and covariates are fitted for each level of the factor, but other parameters are separate.
5.	MODELSTRUCTURE = *parallel*	A separate intercept is fitted for each level of the factor, but other parameters are common.
6.	MODELSTRUCTURE = *null*	A common polynomial trend plus covariate model is fitted, ignoring the factor.

The PRINT option controls the output for these models. The variogram and covariance and correlation matrices can be inspected. The fitting of each model by FITNONLINEAR may be monitored. By default, parameter estimates and standard errors are printed for each model requested. The correlation matrix for the polynomial and covariate parameters may be printed. If a more complex model than MODELSTRUCTURE=full is requested, a summary likelihood analysis table may be printed. Parameter estimates can be SAVEd into a pointer.

```
RMBAL      [PRINT=variogram; GROUP=digtreat; VARIOGRAM=before; \
           YTITLE='Log Weight'; GRAPH=high] y[]; TIMES=0...4

Assessment of within-unit correlation

                u    Variogram(u)   Correlogram(u)
                1        0.001128           0.8677
                2        0.003327           0.6641
                3        0.005214           0.4464
                4        0.006541           0.3399

gamma(0) is              0.009427

Variance-covariance matrix
           1  6.99E-03
           2  7.16E-03  1.09E-02
           3  5.30E-03  8.29E-03  8.61E-03
           4  4.21E-03  5.97E-03  8.22E-03  1.04E-02
           5  2.89E-03  4.22E-03  7.03E-03  9.53E-03  1.03E-02
                    1         2         3         4         5
```

```
Correlation matrix

       1    1.000
       2    0.821    1.000
       3    0.683    0.858    1.000
       4    0.494    0.563    0.870    1.000
       5    0.340    0.399    0.747    0.922    1.000
              1        2        3        4        5

VARIATE    [val=0,1,4,9,16] ts2
RMBAL      [PRINT=e,s;    GROUP=digtreat;    CORR=gauss;    INITIAL=!(.8,.1,.75);    \
           MODEL=full,intercept,covint,null; TREND=1; COVARIATE=!P(ts2); \
           YTITLE='Log Weight'; GRAPH=high] y[]; TIMES=0...4

Fitting full model for digtreat

Maximum likelihood (less constant)
   382.5

Variance model parameter estimates
                                                 estimate        s.e.
phi2  - between-unit variance                      0.7093     0.66620
phi1  - measurement variance                       0.0988     0.07650
beta  - correlation parameter                      0.7640     0.10080
sigma2 - residual variance                         0.0048     0.00150

Trend and covariate estimates
                                                 estimate        s.e.
digtreat 1 constant                                3.9990     0.02246
digtreat 1 linear                                  0.3650     0.01582
digtreat 1 Cov 1                                  -0.0230     0.00365
digtreat 2 constant                                3.9910     0.02928
digtreat 2 linear                                  0.3420     0.02062
digtreat 2 Cov 1                                  -0.0330     0.00476


Fitting intercept model for digtreat

Maximum likelihood (less constant)
   382.4

Variance model parameter estimates
                                                 estimate        s.e.
phi2  - between-unit variance                      0.7209     0.68570
phi1  - measurement variance                       0.1001     0.07880
beta  - correlation parameter                      0.7635     0.10430
sigma2 - residual variance                         0.0048     0.00150

Trend and covariate estimates
                                                 estimate        s.e.
constant                                           3.9962     0.01783
digtreat 1 linear                                  0.3652     0.01580
digtreat 1 Cov 1                                  -0.0234     0.00365
digtreat 2 linear                                  0.3416     0.02059
digtreat 2 Cov 1                                  -0.0329     0.00476
```

```
Fitting covint model for digtreat

Maximum likelihood (less constant)
   372.3

Variance model parameter estimates
                                          estimate        s.e.
phi2  - between-unit variance               0.5467     0.66150
phi1  - measurement variance                0.1141     0.07410
beta  - correlation parameter               0.8325     0.10380
sigma2 - residual variance                  0.0064     0.00230

Trend and covariate estimates
                                          estimate        s.e.
digtreat   constant                         3.9920     0.01971
digtreat 1 linear                           0.3710     0.01671
digtreat 2 linear                           0.3470     0.02179
Cov 1                                       0.3710     0.00302


Fitting null model for digtreat

Maximum likelihood (less constant)
   362.4

Variance model parameter estimates
                                          estimate        s.e.
phi2  - between-unit variance               0.3099     0.46560
phi1  - measurement variance                0.0510     0.02970
beta  - correlation parameter               0.8479     0.05530
sigma2 - residual variance                  0.0109     0.00350

Trend and covariate estimates
                                          estimate        s.e.
constant                                    3.9880     0.02290
linear                                      0.3600     0.01539
Cov 1                                      -0.0270     0.00348


Summary likelihood analysis table

Model                    d.f.         MLL      d.f.      Chi-sq.
full                        6       382.5         *           *
intercept                   5       382.4         1        0.07
covint                      4       372.3         2       20.29
null                        3       362.4         3       40.05
```

RMBAL is limited in that it allows only one factor and no interactions, no variance components in higher strata, and that it uses ML and not REML. Matrix inversion problems have also been found for very highly correlated data. At the moment, RMBAL is useful for its exploratory data analysis and model formulation rather than its capacity for general formal analysis. RMBAL will be superseded in the medium time frame, but will be useful until then, used in conjunction with REML.

References

Box G E P (1950) Problems in the analysis of growth and wear curves *Biometrics* **6** 217-226.
Diggle P J (1990) *Time Series: A Biostatistical Introduction* Clarendon, Oxford.
Kenward M G (1987) A method for comparing profiles of repeated measures *Applied Statistics* **36** 296-308.

Appendix

The procedure code is available from the NAG Genstat web page at:

http://www.nag.co.uk/stats/TT.html

or from the Response Centre at *infodesk@nag.co.uk.*

A procedure for Theil's regression method

M S Dhanoa
Institute of Grassland and Environmental Research
Plas Gogerddan, ABERYSTWYTH
Dyfed, SY23 3EB, UK

1. Introduction

Bivariate linear regression method is one of the most frequently used statistical methods. Most experimenters meet least squares linear regression and have some appreciation of its proper use under the model

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where the ε_i are unobservable random variables independently normally distributed with mean 0 and variance σ^2. The use of least squares linear regression has become so routine that its sensitivity to outliers is not sufficiently appreciated. In biological research and elsewhere, it is inevitable that some of the observations do not conform to the trend suggested by the remaining data values. Many special methods have been developed to overcome problems caused by the outliers. This may involve elimination of apparently rogue values according to some of the regression diagnostics (Cook and Weisberg, 1982), use of weighted least squares, robust estimation (Ross, 1990), etc. In addition, nonparametric or distribution-free methods (Sprent, 1993; Maritz, 1995) have been developed to complement least squares methodology. In the case of poor quality data, these methods can be used just to verify whether solutions derived from least squares method are acceptable or not. If different solutions emerge, then some problem with data under the linear model is indicated and it is prudent to check the data for any anomalies and also the possibility that the linear model is perhaps inappropriate. If necessary, a new model should be defined to account for any nonlinearity. However, if the problem is caused by outliers, then robust or nonparametric regression methods are required. One such distribution-free regression method was proposed by Theil (1950). He suggested that the estimate of regression slope be calculated as the median of the slopes of all lines joining pairs of points with different values of x-variable.

Thus, the purpose of this paper is to implement Theil's regression method in Genstat via an easy to use procedure.

2. Method

Let there be a bivariate sample of size N and also assume that for convenience, data points are arranged in ascending order of the x-variable. Then for any two points (x_i,y_i) and (x_j,y_j), $i<j$, the slope is

$$b_{ij} = (y_j-y_i)/(x_j-x_i)$$

As $b_{ij}=b_{ji}$ for all (i,j), we have $N(N-1)/2$ distinct pairs of points, each leading to a calculated b_{ij}, and these can be arranged in the form of a symmetric matrix

1	*					
2	b_{21}	*				
3	b_{31}	b_{32}	*			
4	b_{41}	b_{42}	b_{43}	*		
⋮						
n	b_{n1}	b_{n2}	b_{n3}	b_{n4}	...	*
	1	2	3	4	...	n

Theil's method states that median (b^*) of all the b_{ij} is the estimate of the slope β and where α is estimated by a^*, where a^* is the median of all $a_i = y_i - b^*x_i$, (although alternatively $\hat{a}$ = median(y_i) - b^*median(x_i) is also

suitable). In the latter case, the fitted line will pass through the median of all observations just like the least squares line passes through the mean. However, use of a^* ensures that the number of data points below and above the Theil fitted line is equal, except when some points lie exactly on the line. If the Theil regression line is very different from the least squares line then it is unwise to use the least squares line because either outliers are present, or the relationship is nonlinear. Hussein and Sprent (1983) found that Theil's method was nearly as efficient as the least squares method when assumptions of normality and homogeneity of variances hold and that it had improved efficiency in the case of long-tail error distributions when sample sizes were <30. They also showed that estimations based on weighted medians were no better in the absence of outliers and were poorer when outliers were present.

3. Confidence Interval

Good estimates a and b of α and β should be such that the residuals e_i, where $e_i=y_i-a-bx_i$, are equally likely to be positive or negative. Sprent (1993) used the fact that

$$b_{ij}=\frac{(y_j-y_i)}{(x_j-x_i)}=\frac{(a+bx_j+e_j-(a+bx_i+e_i))}{(x_j-x_i)}=b+\frac{(e_j-e_i)}{(x_j-x_i)}$$

and thus any b_{ij} is greater than b if (x_i,e_i) and (x_j,e_j) are concordant but less than b if these points are discordant. Critical values of Kendall's tau (τ_c) can be used to calculate the statistic

$$r_c = N(N\text{-}1)(1\text{-}\tau_c)/2.$$

The value of r_c, rounded to the nearest integer, gives the number of smallest ranked, and by symmetry, the highest ranked values of b_{ij} to be rejected to arrive at the $(1\text{-}P_\alpha)$ confidence interval, where P_α is a chosen level of two-tail probability. Due to limited availability of tables of critical values of t_c and programming convenience, the sign test method for setting the confidence interval for a median of a sample was adopted in the Genstat procedure presented here. Theoretically, half the population lies below the median and the other half above it, ignoring any values exactly equal to the median. Thus, the probability of each member of the population being either above or below the median is 0.5. This is a binomial process, and if N observations are drawn then the probability (P) of r members being greater or less than the median is

$$P_r=\binom{n}{r}0.5^n$$

and the cumulative probability is

$$cum\ P_r=\sum_{i=0}^{r}\binom{n}{i}0.5^n$$

Thus, confidence intervals for the median can be calculated using tables of cumulative probability of binomial distribution. Rather than using tables externally or within the procedure, a Genstat procedure BINPROB has been developed, and it is called by the main procedure THEILREG to calculate and print the confidence interval for the Theil slope. For large sample sizes, normal approximation

$$Z=\frac{|r-np-0.5|}{\sqrt{np(1-p)}}$$

where Z is N(0,1), is adequate and the lower confidence limit is calculated as

$$r_c=\frac{N}{2}-\left(Z_{-(1-P_{\alpha/2})}\sqrt{\frac{N}{2}}\right)-\frac{1}{2}$$

where ½ is the continuity correction.

Again we round up to the nearest integer and we reject r_c lowest and r_c highest ranked b_{ij} to arrive at the confidence interval for Theil slope b^*.

4. Jackknife estimate of b^*

It is well known that jackknife can misbehave in the case of order statistics (Efron, 1982; Bissell and Ferguson, 1975), but in the case of the Theil method, the b_{ij} are calculated quantities and their median does not change wildly, even in the presence of an extreme outlier (see later example chosen to show this aspect). Thus far, no misbehaviour has come to light. However, if evidence emerges, then this section of the procedure THEILREG may be expunged. In a sample of size N, (N-1) slopes (b_{ij}) involve each of the individual units. Thus the Nth row and Nth column of the symmetric matrix containing b_{ij} have to be omitted in turn to arrive at N jackknife samples leading to N median estimates of the Theil slope. Because of this complication in declaring subsets of b_{ij}, the Genstat Procedure Library procedure JACKKNIFE could not be used. Therefore, a section is included where jackknife calculations are performed and relevant information is printed out.

5. Procedures

Two procedures are provided. BINPROB calculates the cumulative binomial probability of the r lowest and highest ranked b_{ij} in order to define confidence interval for b^* at pre-selected two-tail probability, e.g., 2.5% in each tail for 95% confidence interval. For $N(N\text{-}1)/2>100$, normal approximation is used. The main procedure THEILREG calculates Theil regression slopes, median slope and confidence interval using BINPROB. This procedure also calculates least squares slope and jackknife refinement of Theil slope. Lineprinter (default) or high resolution graphics are produced including plots using procedure BOXPLOT from the Genstat Procedure Library.

5.1 BINPROB

This procedure calculates cumulative binomial probability when P=0.5 for sample $N(N\text{-}1)/2$, N being the number of units in the input variables. The first parameter is input, n giving $NN1$ (=$N(N\text{-}1)/2$) b_{ij}. The second parameter is output, NTAIL containing the number of r_c lowest and r_c highest ranks which give two tail cumulative probability ≤ 0.005, 0.01, 0.02 and 0.05. The actual cumulative two tail probability of ≤ r_c ranks is saved in the third parameter PTAIL. Both NTAIL and PTAIL are declared as variates of length four before calling BINPROB. Note that this procedure can be used directly for sample size N by preventing it working on Theil sample size $NN1$, by activating the Genstat statement

```
calculate NN1 = N
```

as flagged in the code of the procedure. Results obtained this way check out against available tables (Bradley, 1968; Sprent, 1993).

5.2 THEILREG

The data are passed to the procedure using parameters X_VARIATE and Y_VARIATE. High resolution graphics on the currently active DEVICE, showing boxplot for b_{ij}, data and fitted equations according to Theil method, least squares and jackknifed Theil equation, may be requested by setting option GRAPH=high. Thus, this procedure provides an easy means of using the Theil regression method in day to day applications. In the presence of suspect outliers, the Theil regression method is a useful check for the appropriateness of the least squares line.

6. Example

Data from Example 8.1 in Sprent (1993) is used to illustrate the working of the procedure THEILREG. In this example, one data point is either a rogue value, or a single sample from the range over which the linear trend (defined up to the fifth unit) does not apply. This single value undermines the least squares method but Theil's method gives an equation which is not unduly affected by this outlier. The jackknifed Theil equation appears to be acceptable. A sample calling Genstat program is as follows.

```
open 'binprob.proc' ;ch=3
open 'theilreg.proc' ;ch=4
input [print=p] 3
input [print=p] 4
close 3   :close 4
open 'theil_ink.dat' ;ch=8 ;file=gr
device 8
"Example from Applied Nonparametric Methods (2nd Ed., P Sprent) pp 196"

read [serial=y] x,y
0 1 2 3 4 5 6:
2.5 3.1 3.4 4.0 4.6 5.1 11.1:

THEILREG [graph=high] x;y
stop
```

The output generated by the procedure THEILREG contains the following information:

```
Calculated slopes from Theil regression method
 1       *
 2    0.600       *
 3    0.450   0.300       *
 4    0.500   0.450   0.600       *
 5    0.525   0.500   0.600   0.600       *
 6    0.520   0.500   0.567   0.550   0.500       *
 7    1.433   1.600   1.925   2.367   3.250   6.000       *
         1       2       3       4       5       6       7

The fitted Theil reg. equation is:
   y = 2.333 + 0.5667 x
[The fitted least sq. equation is:
   y = 1.507 + 1.107 x ]

Sample size = 7 ; Theil reg. slopes (N(N-1)/2) = 21

Note:

*** Line 1 => Nominal tail probability levels a (alpha)
*** Line 2 => Number of observations for which cumulative tail Pr<=a
*** Line 3 => Actual probability for the observations in line 2 above
 0.005  0.010  0.025  0.050
    4      4      5      6
0.0036 0.0036 0.0133 0.0392

*** Confidence intervals from the sign test for median ***
90% C. I. for Theil reg. slope ==>    0.5000      <->    0.6000
(Actual probability of the c.i.)     0.9608
95% C. I. for Theil reg. slope ==>    0.5000      <->     1.433
(Actual probability of the c.i.)     0.9867
-----------------------------------------------------

*** Mean of jackknife samples ***
  0.6000   0.6000   0.5500   0.5667   0.5667   0.6000   0.5200
*** Pseudo values ***
  0.3667   0.3667   0.6667   0.5667   0.5667   0.3667   0.8467
*** Jackknifed Theil regression equation ***
   y = 2.459 + 0.5352 x
Bias (slope) =                        0.03143
Standard error (slope) =              0.06926
95% confidence interval (slope)=      (0.3296, 0.7409)
```

In addition to the above information, high resolution graphics on the current DEVICE are available from which hard copies may be obtained. The first segment is boxplot of Theil's b_{ij} and a_i ($=y_i-b^*x_i$), the second contains a boxplot of jackknife samples of b_{ij} and the third has plot of original data with fitted lines according to Theil's method, least squares and jackknifed Theil. Here the calling Genstat program asked for high resolution graphics which are shown in Figure 1(a)-(c).

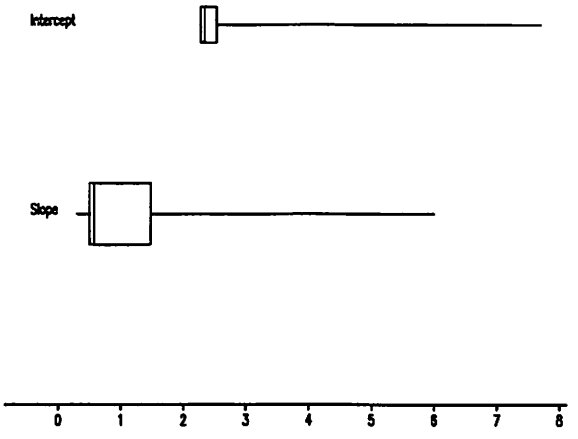


Figure 1(a): Boxplot of Theil regression slopes.

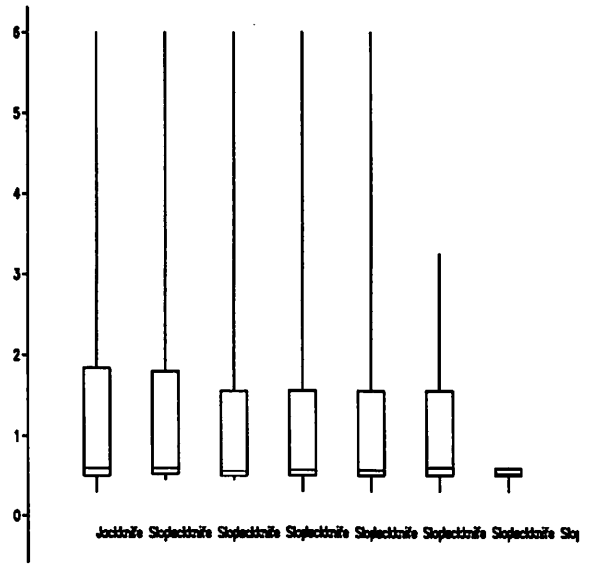


Figure 1(b): Boxplot of jackknife samples of slopes.

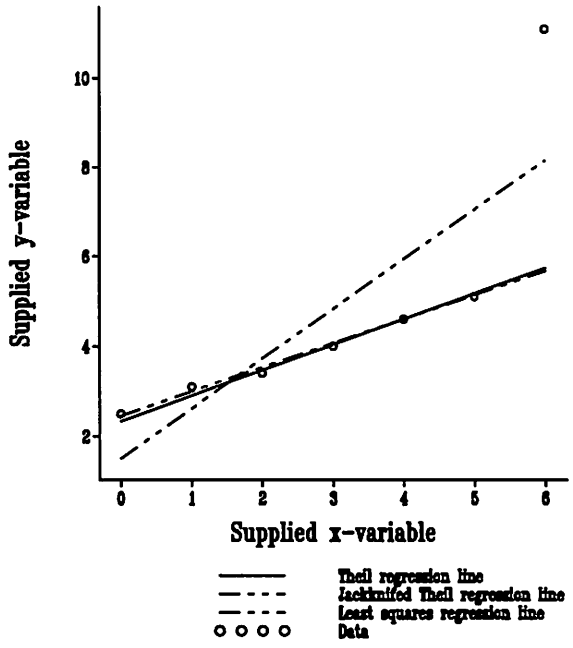


Figure 1(c): Plot of the original data with fitted lines according to the Theil regression method, least squares method and jackknifed Theil regression method.

References

Bissell A F and Ferguson R A (1975) The Jackknife - toy, tool or two-edged weapon *The Statistician* **24** 79-100.

Bradley J V (1968) *Distribution-free Statistical Tests* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Cook R D and Weisberg S (1982) *Residuals and Inference in Regression* Chapman and Hall Ltd., London.

Efron B (1982) *The Jackknife, the Bootstrap and Other Resampling plans* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, USA, 1982.

Hussein S S and Sprent P (1983) Non-parametric Regression *Journal of Royal Statistical Society, (A)* **146** 182-191.

Maritz J S (1995) *Distribution-free Statistical Methods* Chapman and Hall Ltd., London, 1995.

Ross G J S (1990) *Nonlinear Estimation* Springer-Verlag, New York Inc.

Sprent P (1993) *Applied nonparametric Statistical methods 2nd Edition* Chapman and Hall Ltd., London.

Theil H (1950) A rank invariant method of linear and polynomial regression analysis, I, II, III. *Proc. Kon. Nederl. Akad. Wetensch. A,* **53** 386-92, 521-5, 1397-1412.

Appendix

The procedure code is available from the NAG Genstat Web page at:

http://extweb.nag.co.uk/stats/TT.html

or from the Response Centre at *infodesk@nag.co.uk*.

Hit-target radiobiological models

A Kinsella
Department of Mathematics, Statistics and Computer Science
Dublin Institute of Technology
Kevin Street
DUBLIN 8, Ireland

1. Introduction

While the precise detail of the effect of exposing a living organism to a radioactive material is not fully understood, a series of models based on necessarily simplifying assumptions are often found to give reasonable representations of the dose-response relationships observed when studying relatively simple systems such as macromolecules, bacterial and yeast suspensions, and cultured mammalian cells. These materials are used in exploratory studies of drug and/or radiation therapies where the effects of fluctuations due to non-homogeneous populations are reduced by working with large numbers of a single genotype.

The two basic assumptions which underlie the hit-target biological models are (Elkind and Whitmore, 1967):

a) that within a cell there is a *target*, a radio-sensitive volume, which is affected by exposure to a radioactive source, and
b) that the radioactive dose can be modelled as a random process such that every part of the cell surface which is exposed to the radiation source has a constant probability of being *hit* by a radioactive particle.

The addition and multiplication laws of probability, in combination with basic assumptions similar to those underlying the Poisson model in microbiology, give the following dose-response function:

$$\text{Prob(Cell survives radiation dose } d) = \text{Prob(Target not hit in } d \text{ emissions)} = \exp(-d/\delta_0)$$

where the parameter δ_0, the Mean Lethal Dose, is proportional to the probability that a target is hit and inactivated by any single radioactive emission.

This model, the single-target survival model, is immediately extended to the case where n (>1) targets are postulated to exist, the survival function of the multi-target model being

$$\text{Prob(Cell survives radiation dose } d) = 1-(1-\exp(-d/\delta_0))^n,$$

which has an asymmetrical reverse sigmoidal shape. The assumption that a single target must be hit n times to produce the end point gives rise to the same model and when n=1, this model reduces to the single-target form.

A two component model in which there is a both a single-target and a multi-target element has been proposed by Hall (1975), to cater for the situation where the dose-effect curve has a non-zero initial slope. This has implications for radiation protection due to the absence of a lower threshold dose at which biologically significant effects could occur. In addition, this model has implications for the administration of radiation therapy in the treatment of cancer. The survival function is

$$\text{Prob(Cell survives radiation dose } d) = [1-(1-\exp(-d/\delta_0))^n]\exp(-d/\delta_1)$$

which reduces to the single-target form when n=0, and approaches the multi-target form as $\delta_1 \to \infty$.

2. Application

The data shown below are results from an invitro study of the radio-protective properties of a pharmaceutical product. Treated cells were plated in triplicate, irradiated with medium energy X-rays and, after incubation, the numbers of surviving cells counted by direct inspection of the cell culture plates. The cell concentrations which were initially plated, increase with dose and were chosen to yield reasonable post–irradiation densities.

Table 1: Surviving cell counts

Dose	0	196	326	457	587	782	848	979	1110
Cell Concentration	101	167	258	513	696	2565	4882	11901	19423
Counts for replicate 1	51	59	49	60	40	54	74	89	54
Counts for replicate 2	101	64	36	58	53	57	57	94	68
Counts for replicate 3	76	59	50	58	45	40	65	95	60

The individual surviving cell counts are postulated to conform to a Poisson model and this model is compounded with the cell survival models, to produce Poisson models with the following expected values:

Single–target	$E(y_{ij}) = c_{ij}\exp(-d/\delta_0)$
Multi-target	$E(y_{ij}) = \mu c_{ij}[1-(1-\exp(-d/\delta_0))^n]$
Modified Multi–target	$E(y_{ij}) = \mu c_{ij}[\ 1-(1-\exp(-d/\delta_0))^n\ ]\exp(-d/\delta_1)$

Here y_{ij} represents the survival count of the jth replicate (j=1,2,3) at the ith dose (i=1,...,9), μ denotes the zero dose count, i.e., the natural survival level, and c_{ij} denotes the cell concentration of the jth replicate plated at the ith dose.

Since the models are linear in the parameter μ and nonlinear in the parameters δ_0, δ_1 and n, the FITNONLINEAR directive was used to estimate the parameters of the models in the following manner for the single-target model:

```
CALCULATE D = D/100
EXPRESSION e1; VALUE = !e(Z = CC*(EXP(-D/D0)))
MODEL [ DISTRIBUTION = poisson ] X
RCYCLE D0; INITIAL = 2.0
FITNONLINEAR [ PRINT = summary,estimates,fittedvalues;\
  CALCULATION = e1; CONSTANT = omit; SELINEAR = yes ] Z
```

The dose variate is denoted by D which is scaled to ensure that the D0 parameter is of the same magnitude as the other model parameters. The cell concentration variate is denoted by CC while the response variate, the surviving cell count, is denoted by X.

In the case of the multi-target model VALUE is altered to

```
VALUE = !e(Z = CC*(1 - (1-EXP(D/D0))**NU))
```

and RCYCLE becomes

```
RCYCLE = D0, NU ; INITIAL = 2.0, 1.0
```

while in the case of the modified multi–target model the corresponding alterations are

```
VALUE = !e(Z = CC*(1 - (1-EXP(-D/D0))**NU) * EXP(-D/D1))
RCYCLE = D0, D1,NU ; INITIAL = 3.0, 3.0, 3.0
```

The parameter estimates, standard estimates and residual deviances for each model are given in Table 2.

Table 2: Parameter Estimates

Model	$\hat{\mu}_0$	δ_0	$\hat{n}$	δ_1	Residual Deviance	df
Single-target	0.878	2.056			59.10	25
s.e.	0.0401	0.0269				
Multi-target	0.742	1.933	1.556		44.86	24
s.e.	0.0488	0.0392	0.1870			
Modified Multi-target	0.754	4.983	3.480	2.611	42.55	23
s.e.	0.0495	0.9400	2.4100	0.4670		

Using the reduction in the residual deviances as a measure of model fit, the multi-target model is a considerable improvement on the single-target model, while there is little advantage in choosing the four parameter modified multi-target model over the three parameter multi-target model.

The largest standardised residual (-2.94) corresponds to the observed count of 101 at the 0 dose, the multi-target model predicted value being 88.71. The model is extended to cater for this extreme value by including an extra term with corresponding dummy data variate X1, which takes value 1 when X=101 and 0 elsewhere. The modifications to VALUE and RCYCLE are

```
VALUE = !e (Z = CC*(1 - (1-EXP(-D/D0))**NU) + CC*B1*X1)
RCYCLE = D0,D1,NU,B1; INITIAL = 2.0,2.0,2.0,1.0
```

The largest standardised residual which was found after fitting this model was +2.36, corresponding to the 53 count at dose 696, the predicted value being 37.63. Extending the model in similar fashion and including a second dummy data variate X2, involves using the following modified VALUE and RCYCLE:

```
VALUE = !e (Z = CC *(1 - (1-EXP(-D/D0))**NU) + CC*B1*X1 + CC*B2*X2)
RCYCLE = D0,D1,NU,B1,B2; 2.0,2.0,2.0,1.0,1.0
```

The parameter estimates, standard errors and residual deviances for these extended versions of the multi-target model are given in Table 3.

Table 3: Parameter estimates, Extended Multi–target Models

Model	$\hat{\mu}_0$	δ_0	$\hat{n}$	β_0	β_1	Residual Deviance	df
Removing 101 count	0.860	1.951	1.283	-0.413		32.63	23
s.e.	0.0932	0.0390	0.1600	0.0932			
Removing both points	0.865	1.962	1.234	-0.416	0.027	26.82	22
s.e.	0.0646	0.0396	0.1550	0.0926	0.0124		

References

Elkind M M and Whitmore G F (1967) *The Radiobiology of Cultured Mammalian Cells* Gordon and Breach, London.

Hall E J (1975) Biological problems in the Measurement of survival at low dose, *Cell Survival after Low Doses of Radiation: Theoretical and Clinical Implications* Wiley, London.

Hierarchical generalized linear models with up to five random components

J A Nelder
Department of Mathematics
Imperial College
LONDON SW7 2BZ, UK

j.nelder@thor.ma.ic.ac.uk

HGLMs are a generalization of generalized linear mixed models (GLMMs), in which the additional random components are not restricted to be Normal. In particular, useful properties result if the conjugate distribution is allowed for these components. The REML approach for Normal models generalizes to this new class for the estimation of dispersion components.

The procedures in the file **pr.hg** allow the fitting of seven HGLMs. They are designed for Release 5.3.2 of Genstat and require the K-system code for that release. Of the seven HGLMs, three are conjugate HGLMs, Poisson-Gamma, binomial-beta and gamma-inverse gamma; three are GLMMs, Poisson-Normal, binomial-Normal, and gamma-Normal, while the seventh is the classical Normal-Normal. Up to five random effects u are allowed. The first distribution in each name is that of y, the response, given u, and the second is that of u. The systematic part of the model is of the GLM form:

$$\eta = X\beta = g(\mu).$$

1. Conjugate HGLMs

For three random components u_1, u_2 and u_3, we have the specification

$E(y \mid u) = \mu'$ where $\eta' = g(\mu')$
$\eta' = \eta + v_1 + v_2 + v_3$, where $v_1 = v_1(u_1)$ etc.
y has a GLM distribution and u has the conjugate distribution

For other numbers of random components, use the corresponding numbers of u and v functions.

The four supported models have the following distributions and functions :

	y distribution	u distribution	$g()$	$v()$
Poisson-gamma	Poisson	gamma	log	log
binomial-beta	binomial	beta	logit	logit
gamma-inv.gamma	gamma	inverse gamma	log	log
Normal-Normal	Normal	Normal	identity	identity

2. GLMMs

For these models the second distribution is Normal and $v(u)=u$. Again the first distribution may be Poisson, binomial, or gamma.

3. Dispersion parameters

The first distribution has one parameter to be estimated if Normal (the variance σ_0^2) or gamma (the shape parameter ν). For one random component, the second distribution has one parameter (α_1) if gamma, two (α_1, α_2)

if beta, one (α, the shape parameter) if inverse-gamma, and one (σ_1^2, the variance) if Normal. For two random components the names are (α_{11}, α_{21}) if gamma, (α_{11}, α_{21} ; α_{12}, α_{22}) if beta, (α_1,α_2) if inverse-gamma, and (σ_1^2, σ_2^2) if Normal. For three random components the corresponding names are (α_{11}, α_{21}, α_{31}), (α_{11}, α_{21} ; α_{12},α_{22}; α_{13}, α_{23}), (α_1, α_2, α_3) and (σ_1^2, σ_2^2, σ_3^2); and so on.

4. Criteria for estimation

For given values of the dispersion components, the algorithm maximizes the hierarchical (*h*-) likelihood to estimate β and u_i. The log-*h* likelihood l is defined with, for example, three random components as

$$l = l(\beta \mid y,u) + l(u_1) + l(u_2) + l(u_3)$$

This is not a joint likelihood in the orthodox sense because the u are not observable. Note that use of *h*-likelihood avoids the integration required for the marginal-likelihood approach, while at the same time giving estimates of the random components u.

For given values of β and u, the dispersion components are estimated by maximising a generalization of the REML criterion for Normal-Normal models. This is an adjusted profile *h*-likelihood, using the formulation of Cox and Reid.

5. Fitting an HGLM model

There are ten procedures, all ending in hg, which are described in Section 6. These procedures require the use of the K-system. For initial installation of the hg system, see Section 8. To use the system with a data set, set up for analysis as follows:

(i) read in the data
(ii) if an offset is required, set it in `fmoff`
(iii) to set a prior weight w, use the K-system statement `wei w`
(v) set the model for fixed effects as a formula in `fmod`
(vi) set the number of units and the response variate using the K-system statements

```
kun <no. of units> : yvar <y>
```

(vii) for the binomial-beta and binomial-Normal models, set the binomial denominator with the K-system statement `err b; <bd>`

Set up the model with `suhg` and initial estimates of the dispersion parameters with `sdhg`. To fit, alternate `fhg` and `edhg` as required.

NB: Do not omit points from the data by setting a prior weight containing zeros, as this may cause the degrees of freedom in `edhg` to be wrongly set. Instead use the library procedure `SUBSET`, with the option `NLEVS` set to `yes`, to compact the data.

6. Checking an HGLM model

After fitting a model, the model-checking facilities of the K-system may be used. In addition, a check on the estimated random effects may be made by invoking `knpl dresu[1...nrc]`, where `nrc` is the number of random components. The vectors `dresu[]` contain the deviance residuals from a null GLM, fitted to the estimated random effects `u[]`. The plots should show approximately linear trends if the model is satisfactory.

7. The procedures

`vershg` Summarizes the state of the current version

`helphg` Gives summary of procedures and their parameters.

`hgmod` Gives a summary of the current hglm.

`dehg {[<integer>]}` Displays estimates etc. of fixed effects. Number of decimal places is abs(integer); negative sign gives E–format.

`suhg {[<random components>]} <letter1>;<letter2> {;<formula>}`
Sets up model with possible combinations `p;g p;n b;b b;n g;i g;n n;n` for `letter1` and `letter2`. Options define random components, separated by semi-colons. Components may be intercepts (denoted by factors) or slopes (denoted by variates). The first element must be a factor. A random slope is connected to the nearest previous factor. Thus in the specification

```
f1; v1; v2; f2; v3
```

the random slopes defined by `v1` and `v2` are connected with intercepts defined by factor `f1` and those of `v3` with factor `f2`.

`sdhg <number0>; <numbers1...nrc>`
Sets initial values of dispersion parameters, separated by semi-colons. If first distribution is Poisson or binomial, `<number0>` (but not the following semi-colon) may be omitted. If the first distribution is gamma, set `<number0>` to the shape parameter, and if Normal, to the variance. For random components with gamma distributions `<numbers1...nrc>` are values for shape parameters; if beta, `<numbers1...nrc>` are pairs of values (separated by commas) for the two parameters (a symmetric distribution needs one value only). For the inverse gamma, the numbers are shape parameters of the distribution. For Normal components, values of variances are required.

`fhg{[<scalar>; <letter>; <letter> ]} {newy}`
Fits β and u for HGLM as set in `suhg` with current estimates of dispersion parameters as set by `sdhg` or `edhg`. The first option defines the number of cycles in the iteration (default 3); the second (`NULL`) defines the fixed-effect model to be null (i.e. without even an intercept) with setting `y`, else non-null `n` (default); the third option is the `CONSTANT` option of `FIT` (default `e`). The optional parameter, if set to `newy`, allows fitting with a new y-variate without recalculation of the augmented data matrix. Note that if `fmod` or the `NULL` option is reset, the model must be respecified, starting with `suhg`.

`aplhg` Calculates and prints the adjusted profile h-likelihood; It is used as a criterion for fitting the dispersion parameters.

`edhg {[<letter>]}` Estimates dispersion parameters given estimates of β and u, using one cycle of Newton method applied to logs of the components. The option allows the α parameters in the binomial-beta model to be kept equal (`e`) or allowed to be unequal (`u`). Default `e`.

`clearhg` Clears existing fit if another data set is to be loaded. Note that if the two data sets have common identifiers these should be deleted separately using `kdel` before the new data set is loaded.

8. The hg system files

These are:

in.hg	Initializes the common blocks, expressions, etc.
pr.hg	Holds the procedures for HGLM analysis.
mk.hg	Makes the system as a backing-store file.
su.hg	Sets up the system on subsequent runs.
mn.hg	The system manual.

The files are assumed to be stored in gen532\ksys. To make the system initially, or after the changes, type

```
g532  in2=gen532\ksys\mk.hg s=30
```

then type

```
input 2
```

For subsequent use, replace **mk.hg** with **su.hg**. In either case the K-system procedures will be loaded automatically.

If necessary, change **g532** to your name for version 5.3.2; the space setting s=30 can be changed as required.

Note that **in.hg** contains details of the underlying data structures that occur in the workspace used by the system.

All system files, together with the K-system which they use, are available from the NAG Web page at http://www.nag.co.uk/stats/TT.html. Please report any bugs found, or difficulties with the documentation. The models are described and illustrated in Lee and Nelder (1996).

References

Lee Y and Nelder J A (1996) Hierarchical Generalized Linear Models *Journal of the Royal Statistical Society Series B* **58** 619-678.

Editor's Note

Some of the procedures mentioned above have been updated since this article was written. The most recent version can be found via the NAG/Genstat Web page at:

http://www.nag.co.uk/stats/TT.html

Using Genstat to screen covariates in mixed linear models

A M Richardson
School of Mathematics and Statistics
Faculty of Information Sciences and Engineering
University of Canberra
PO Box 1
Belconnen ACT 2616
Australia.

1. Introduction

This paper opens with discussion of a real data set which will be used to illustrate the methods discussed. Notation for the mixed linear model is then introduced. This is followed by discussion of two types of likelihood ratio tests: one based upon changing the model and one based upon changing the likelihood. The change of deviance test and Wald test are then discussed, along with two other methods of covariate screening. The paper concludes with results from applying a number of these test statistics to a real data set in Genstat, and a short summary.

2. The mixed linear model

The data used here as an illustration are from a crop variety trial in Scotland, reported by Patterson and Nabugoomu (1992). Six varieties of wheat were grown at ten centres that formed a sample of the main types of growing area for wheat in Scotland, and the yields in tonnes/hectare were recorded. The experiment is unbalanced because of 60 possible variety-centre combinations, only 46 were used. At three centres all six varieties were grown but at the other seven, only four varieties were grown.

A mixed linear model to describe the wheat data is of the form

$$y_{ij} = \text{mean} + \text{variety}_i + \text{centre}_j + \text{error}_{ij}, \quad i = 1, ..., 6; j = 1, ..., 10.$$

This can be written in matrix form as

$$y = X\alpha + Z\beta + \epsilon,$$

where y is n by 1 ($n = 46$) and can be treated as consisting of 10 independent subvectors, corresponding to the 10 centres i.e., the random part of the model is nested. The vector α is p by 1 ($p = 6$); β is g by 1 ($g = 10$). Further, $\mathrm{E}(\beta) = 0$, $\mathrm{var}(\beta) = \sigma^2 I_g$, $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2 I_n$ and β and ϵ are assumed to be independent. Thus $\mathrm{E}(y) = X\alpha$ and $\mathrm{Var}(y) = \sigma^2_\beta ZZ' + \sigma^2_\epsilon I_n = V$.

In the case of the wheat data,

$$\begin{bmatrix} 1_4 & & & & \\ & 1_4 & & & \\ & & \ddots & & \\ & & & 1_6 & \\ & & & & 1_4 \end{bmatrix}$$

where 1_m denotes an m-vector of ones. The structure of Z leads to a block diagonal structure for V, with 10 blocks. A typical block looks like

$$\begin{bmatrix} \sigma^2_\beta + \sigma^2_\epsilon & & \sigma^2_\beta \\ & \ddots & \\ \sigma^2_\beta & & \sigma^2_\beta + \sigma^2_\epsilon \end{bmatrix}.$$

The parameters requiring estimation are α, σ^2_β and σ^2_ϵ. To ensure identifiability of the parameters in the wheat data, we will set $\alpha_1 = 0$.

This paper focuses on covariate screening, that is, testing whether certain covariates can be excluded from the model. Without loss of generality, we will test that the first s covariates can be omitted. In this case, the null hypothesis can be written in matrix form as

$$H_0 : H'\alpha = 0$$

where $H' = (I_s : 0)$ and $s \leq p$. The matrix H consists of an $s \times s$ identity matrix next to a $(p-s) \times s$ matrix of zeroes.

3. Likelihood ratio tests: changing the model

We will denote the full model by $y = X_F\alpha_F + Z\beta + \epsilon$ and the reduced model by $y = X_R\alpha_R + Z\beta + \epsilon$. If β and ϵ are normally distributed, then the log likelihood for a general model with fixed-effects matrix X is

$$l(\tau) = (-1/2)\{(y - X\alpha)'V^{-1}(y - X\alpha) + \log|V|\}.$$

The likelihood ratio test statistic for testing H_0 is

$$\begin{aligned} 2[l(\hat{\tau}_F) - l(\hat{\tau}_R)] =& (y - X_R\hat{\alpha}_R)'\hat{V}_R^{-1}(y - X_R\hat{\alpha}_R) + \log|\hat{V}_R| \\ & -(y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) - \log|\hat{V}_F|. \end{aligned}$$

This test statistic has an approximate χ^2_s distribution under normality and H_0.

The small-sample bias in maximum likelihood estimates of variance parameters has, however, led to a trend towards estimation techniques that correct for bias, such as restricted maximum likelihood (see Harville (1977)). The restricted log likelihood can be derived in a number of ways: by projecting the data onto the residual space, by constructing the likelihood of error contrasts, by conditional or by Bayesian arguments. The end result is the following log-likelihood

$$\lambda_R(\theta) = (-1/2)\{(y - X\alpha(\theta))'V^{-1}(y - X\alpha(\theta)) + \log|V| + \log|X'V^{-1}X|\},$$

where $\alpha(\theta) = (X'V^{-1}X)^{-1}X'V^{-1}y$.

This log-likelihood is a function of θ alone and as such cannot be used to test hypotheses about α. However, from the form of $\lambda(\theta)$, it is clear that α has been concentrated out, and it is simple to obtain the following unconcentrated log-likelihood

$$\lambda^*(\tau) = (-1/2)\{(y - X\alpha)'V^{-1}(y - X\alpha) + \log|V| + \log|X'V^{-1}X|\}$$

which leads to a candidate test statistic

$$\begin{aligned} 2[\lambda^*(\hat{\tau}_F) - \lambda^*(\hat{\tau}_R)] =& (y - X_R\hat{\alpha}_R)'\hat{V}_R^{-1}(y - X_R\hat{\alpha}_R) + \log|\hat{V}_R| + \log|X_R'\hat{V}_R^{-1}X_R| \\ & -(y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) - \log|\hat{V}_F| - \log|X_F'\hat{V}_F^{-1}X_F|. \end{aligned}$$

However, this statistic cannot be used as a test statistic because it is not scale invariant. If X consists of columns of zeroes and ones, then the likelihood ratio based on ky for any constant k differs from that based on y by $2s\log(k)$. This means that certain choices of k can lead to negative "likelihood ratios".

The expression that Genstat (1993) refers to as the restricted log-likelihood comes from Harville (1974) and looks like

$$RL(\theta) = (-1/2)\{(y - X\alpha(\theta))'V^{-1}(y - X\alpha(\theta)) + \log|V| + \log|X'V^{-1}X| - \log|X'X| + (n-p)\log 2\pi\}$$

which in unconcentrated form looks like

$$RL^*(\theta) = (-1/2)\{(y - X\alpha)'V^{-1}(y - X\alpha) + \log|V| + \log|X'V^{-1}X| - \log|X'X| + (n-p)\log 2\pi\}.$$

A candidate for a likelihood ratio test statistic would then be

$$\begin{aligned} 2[RL^*(\hat{\tau}_R) - RL^*(\hat{\tau}_F)] = & (y - X_R\hat{\alpha}_R)'\hat{V}_R^{-1}(y - X_R\hat{\alpha}_R) + \log|\hat{V}_R| + \log|X_R'\hat{V}_R^{-1}X_R| \\ & - (y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) - \log|\hat{V}_F| - \log|X_F'\hat{V}_F^{-1}X_F| \\ & - \log|X_R'X_R| + \log|X_F'X_F| + s\log 2\pi \end{aligned}$$

but this statistic is not scale invariant either.

4. Likelihood Ratio Tests: Changing the Likelihood

Genstat (1993) describes the likelihood ratio test in Genstat 5 Release 3.1 in these terms. "The statistic proposed by Welham and Thompson can be used to test a fixed model against a nested submodel. The method calculates the likelihood for the full fixed model as usual. The same projection is then used for the submodel and fixed effects to be dropped in the submodel are constrained to be zero. The difference in log-likelihoods therefore gives a likelihood ratio test in the usual way...". Using Welham and Thompson (1997), it can be seen that the restricted log-likelihood ratio test in Genstat 5 Release 3.1 is defined by taking twice the difference of the following likelihoods.

Restricted likelihood under full model:

$$RL_F(\theta) = (-1/2)\{(y-X_F\alpha_F(\theta))'V_F^{-1}(y-X_F\alpha_F(\theta))+\log|V_F|+\log|X_F'V_F^{-1}X_F|-\log|X_F'X_F|+(n-p)\log(2\pi)\}$$

where $\alpha_F(\theta) = (X_F V_F^{-1} X_F)^{-1} X_F V_F^{-1} y$.

Restricted likelihood under reduced model:

$$RL_R(\theta) = (-1/2)\{(y-X_R\alpha_R(\theta))'V_R^{-1}(y-X_R\alpha_R(\theta))+\log|V_R|+\log|X_F'V_R^{-1}X_F|-\log|X_F'X_F|+(n-p)\log(2\pi)\}$$

where $\alpha_R(\theta) = (X_R V_R^{-1} X_R)^{-1} X_R V_R^{-1} y$.

Unconcentrated versions of these likelihoods are easily obtained, as follows:

$$RL_F^*(\tau) = (-1/2)\{(y-X_F\alpha_F)'V_F^{-1}(y-X_F\alpha_F)+\log|V_F|+\log|X_F'V_F^{-1}X_F|-\log|X_F'X_F|+(n-p)\log(2\pi)\}$$

$$RL_R^*(\tau) = (-1/2)\{(y-X_R\alpha_R)'V_R^{-1}(y-X_R\alpha_R)+\log|V_R|+\log|X_F'V_R^{-1}X_F|-\log|X_F'X_F|+(n-p)\log(2\pi)\}.$$

A test statistic can be constructed in the usual way

$$\begin{aligned} 2\log(LR) &= 2[RL_F^*(\hat{\tau}_F) - [RL_R^*(\hat{\tau}_R)] \\ &= (y - X_R\hat{\alpha}_R)'\hat{V}_R^{-1}(y - X_R\hat{\alpha}_R) + \log|\hat{V}_R| + \log|X_F'\hat{V}_R^{-1}X_F| \\ &\quad - (y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) - \log|\hat{V}_F| - \log|X_F'\hat{V}_F^{-1}X_F|. \end{aligned}$$

The statistic is scale invariant, has an approximately χ_s^2 distribution under normality and H_0, and is available in Genstat output. However, because the estimating equations for these two likelihoods are different, the variance components must be re-estimated for each sub-model. This fact is not made clear in Genstat (1993), but is discussed more fully in Welham and Thompson (1997).

It is also worth noting that for the regression model and maximum likelihood estimation, likelihood ratio tests constructed by changing the model and by changing the likelihood are identical. For the regression model and REML estimation, the two approaches differ but the resulting test statistics are both scale invariant.

5. Change of deviance tests

Two key features of the construction of the change of deviance test proposed here are worthy of mention. Firstly, the deviance is specified by the procedure used to estimate the fixed effects, namely weighted least squares. It is quite clear from the preceding examples of test statistics (some that have good properties, some that do not) that the weighted least squares criterion is quite stable, and it is usually other terms in the likelihood for all the parameters that cause an otherwise useful-looking statistic to have undesirable properties. Secondly, the covariance matrix is held fixed across different models for the fixed effects. This has the effect of significantly decreasing the computational burden, compared to the two likelihood ratio test statistics described above.

Therefore, the change of deviance statistic we propose is

$$2\Delta = (y - X_R\hat{\alpha}_R)'\hat{V}_F^{-1}(y - X_R\hat{\alpha}_R) - (y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F).$$

The variance components $\hat{\theta}$ and hence $\hat{V}$ are found using REML and the full model.

The fixed effects $\hat{\alpha}$ are the weighted least squares estimates of α with estimated weights. Weighted least squares estimates with estimated weights minimise the following expression:

$$Q(\alpha) = (1/2)(y - X\alpha)'\hat{V}^{-1}(y - X\alpha)$$

which can be expresssed equivalently as the roots of the equation

$$\partial Q(\alpha)/\partial\alpha = X'\hat{V}^{-1}(y - X\alpha) = 0.$$

The solution is of course well-known:

$$\alpha(\theta) = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y.$$

In order to give the asymptotic distribution of the change of deviance statistic, we need to define two matrices. They are

$$G_{\alpha\alpha} = -E[\partial^2 Q(\alpha)/\partial\alpha\partial\alpha']$$

and

$$F_{\alpha\alpha} = E[(\partial Q(\alpha)/\partial\alpha)(\partial Q(\alpha)/\partial\alpha)'].$$

It can then be shown that $2\Delta D \sim \Sigma_{i=1}^{s} w_i U_i$, where $U_i \sim \chi_1^2, 1 \leq i \leq s$, and w_i are the eigenvalues of the matrix $[G_{\alpha\alpha}^{11}]^{-1}[G_{\alpha\alpha}^{-1}F_{\alpha\alpha}G_{\alpha\alpha}^{-1}]_{11}$. The subscript 11 denotes the top left $s \times s$ submatrix of $G_{\alpha\alpha}^{-1}F_{\alpha\alpha}G_{\alpha\alpha}^{-1}$ and the superscript 11 denotes the top left $s \times s$ submatrix of $G_{\alpha\alpha}^{-1}$.

The distribution of $2\Delta D$ simplifies if $G_{\alpha\alpha} = wF_{\alpha\alpha}$; in other words, if $F_{\alpha\alpha}$ and $G_{\alpha\alpha}$ are a scalar multiple of each other. In that case $2w\Delta D \sim \chi_s^2$.

For the normal model assumed earlier, this simplification applies. In fact, $G_{\alpha\alpha} = F_{\alpha\alpha} = X'V^{-1}X$ and w_i turn out to be the eigenvalues of $[G_{\alpha\alpha}^{11}]^{-1}[G_{\alpha\alpha}^{-1}]_{11} = I_s$. Then $2\Delta D \sim \chi_s^2$.

The change of deviance statistic is referred to in Genstat (1993) in the following terms. "Other proposals have been made for the testing of fixed effects using REML estimation procedures. Several of these are based on estimating the full fixed model, fixing the values of the gammas [variance components], and then estimating the nested sub-model. The change in residual sum of squares under this procedure is [asymptotically] equivalent to the Wald statistic."

In the next section, we will prove that asymptotic equivalence.

6. Wald Statistics

First we define Wald statistics in the context of covariate screening in mixed linear models. The Wald statistic is

$$\begin{aligned} W &= \hat{\alpha}_{F1}'[\text{var}(\hat{\alpha}_{F1})]^{-1}\hat{\alpha}_{F1} \\ &= \hat{\alpha}_{F1}'[G_{\alpha\alpha}^{11}]^{-1}\hat{\alpha}_{F1}. \end{aligned}$$

The Wald statistic has an asymptotic χ_s^2 distribution under normality and H_0.

To show the asymptotic equivalence under normality of the Wald statistic and the change of deviance statistic asserted by Genstat (1993), we note first under the full model and given $\hat{\theta}_F$,

$$X_F'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) = 0.$$

Expanding this equation in a Taylor series about (α_0, θ_0), we obtain

$$X_F'\hat{V}_0^{-1}(y - X_F\alpha_0) - G_{\alpha\alpha}(\hat{\alpha}_F - \alpha_0) = 0$$

which can be rearranged to read

$$(\hat{\alpha}_F - \alpha_0) = G_{\alpha\alpha}^{-1} X_F' V_0^{-1} (y - X_F \alpha_0).$$

Before performing a similar expansion under H_0, we will first change notation. Until now, α_R has referred to a vector of length $p - s$; now it will refer to a vector of length p, of which the first s entries are of course zero. Then, using the method of Lagrange multipliers, we note that under H_0,

$$\begin{aligned} 0 &= X_F' \hat{V}_F^{-1} (y - X_F \hat{\alpha}_R) - H\eta \\ 0 &= H' \hat{\alpha}_R \end{aligned}$$

where η is the Lagrange multiplier. This pair of equations can be expanded in a Taylor series to obtain

$$(\hat{\alpha}_R - \alpha_0) = \{G_{\alpha\alpha}^{-1} - G_{\alpha\alpha}^{-1} H [H' G_{\alpha\alpha}^{-1} H]^{-1} H' G_{\alpha\alpha}^{-1}\} X_F' V^{-1} (y - X_F \alpha_0).$$

Subtracting one Taylor expansion from the other we obtain

$$(\hat{\alpha}_F - \hat{\alpha}_R) = \{G_{\alpha\alpha}^{-1} H [H' G_{\alpha\alpha}^{-1} H]^{-1} H' G_{\alpha\alpha}^{-1}\} X_F' V_0^{-1} (y - X_F \alpha_0).$$

Now the change of deviance statistic is

$$2\Delta D = (y - X_F \hat{\alpha}_R)' \hat{V}_F^{-1} (y - X_F \hat{\alpha}_R) - (y - X_F \hat{\alpha}_F)' \hat{V}_F^{-1} (y - X_F \hat{\alpha}_F)$$

Expanding the first term in a two-term Taylor series about $\hat{\alpha}_F$ and tidying up we obtain

$$2\Delta D = (y - X_F \alpha_0)' V_0^{-1} X_F G_{\alpha\alpha}^{-1} H [H' G_{\alpha\alpha}^{-1} H]^{-1} H' G_{\alpha\alpha}^{-1} X_F' V_0^{-1} (y - X_F \alpha_0)$$

However, the Wald statistic is

$$\begin{aligned} W &= \hat{\alpha}_{F1}' [G_{\alpha\alpha}^{11}]^{-1} \hat{\alpha}_{F1} \\ &= (\hat{\alpha}_{F1} - \hat{\alpha}_{R1})' [H' G_{\alpha\alpha}^{-1} H]^{-1} (\hat{\alpha}_{F1} - \hat{\alpha}_{R1}) \\ &= (\hat{\alpha}_F - \hat{\alpha}_R)' H [H' G_{\alpha\alpha}^{-1} H]^{-1} H' (\hat{\alpha}_F - \hat{\alpha}_R) \end{aligned}$$

Substituting for $(\hat{\alpha}_F - \hat{\alpha}_R)$ and tidying up, we see that

$$W = (y - X_F \alpha_0)' V_0^{-1} X_F G_{\alpha\alpha}^{-1} H [H' G_{\alpha\alpha}^{-1} H]^{-1} H' G_{\alpha\alpha}^{-1} X_F' V_0^{-1} (y - X_F \alpha_0)$$

which implies that $2\Delta D = W$ as required.

7. Other Methods of Covariate Screening

We will mention two other possibilities here: score tests and robust tests.

The score test is based on the estimating equation for α, which can be thought of as the derivative of the normal likelihood with respect to α. The score test statistic is as follows:

$$S = [X_F' \hat{V}_R^{-1} (y - X_F \hat{\alpha}_R)]_1' [G_{\alpha\alpha}^{11}] [X_F' \hat{V}_R^{-1} (y - X_F \hat{\alpha}_R)]_1$$

As before, under normality and H_0, S has an asymptotic χ^2 distribution on s degrees of freedom. The approximation is often better than for the Wald test, because the score is closer to being a sum of independent normally distributed random variables than a parameter estimate is. Furthermore, it can be shown, using Taylor expansions, that the score test is asymptotically equivalent to the Wald test and hence to the change of deviance statistic.

We will only give a brief indication of test statistics for robust covariate screening. These are a generalisation of the change of deviance test. The proposed test statistic is

$$2\Delta GD = 2[\rho(\tilde{V}_F^{-1/2} (y - X_F \tilde{\alpha}_F)) - \rho(\tilde{V}_F^{-1/2} (y - X_R \tilde{\alpha}_R))]$$

where $\tilde{V}$ and $\tilde{\alpha}$ are robust estimates of the parameters. Here we have returned to the notation that treats α_R as having length $p - s$.

The asymptotic distribution of $2\Delta GD$ is the same as $\Sigma_{i=1}^{s} w_i U_i$, where $U_i \sim \chi_1^2$, $1 \leq i \leq s$ and w_i are the eigenvalues of the matrix $[G_{\alpha\alpha}^{-1}]_{11}^{-1}[G_{\alpha\alpha}^{-1}T_{\alpha\alpha}G_{\alpha\alpha}^{-1}]_{11}$. The matrices involved are defined as follows:

$$G_{\alpha\alpha} = -E[\partial^2\rho(\alpha)/\partial\alpha\partial\alpha']$$
$$F_{\alpha\alpha} = E[(\partial\rho(\alpha)/\partial\alpha)(\partial\rho(\alpha)/\partial\alpha)']$$
$$T_{\alpha\alpha} = G_{\alpha\alpha}G^{\alpha\alpha}F_{\alpha\alpha}G^{\alpha\alpha}G_{\alpha\alpha} + G_{\alpha\alpha}G^{\alpha\alpha}F_{\alpha\theta}G^{\theta\alpha}G_{\alpha\alpha} + G_{\alpha\theta}G^{\theta\theta}F_{\theta\alpha}G^{\alpha\alpha}G_{\alpha\alpha} + G_{\alpha\theta}G^{\theta\theta}F_{\theta\theta}G^{\theta\alpha}G_{\alpha\alpha}.$$

Robust Wald and score tests have been studied by Heritier and Ronchetti (1994) for the regression model i.e., only one variance component. That work has not yet been extended to mixed linear models. The fact that the distribution of the robust test statistic is a weighted sum of χ_1^2 variables need not cause computational problems since an F approximation exists for such linear combinations.

8. Results: wheat data

Recall that the wheat data involved six varieties of wheat. An obvious hypothesis to test is that of no difference between the six varieties. Using Genstat 5 Release 3.1, we have the REML estimation and three statistics available, namely $2\log(LR), W$ and $2\Delta D$.

Genstat (1993) notes that "the deviance given in the monitoring information differs from the deviance given by `PRINT=deviance`, since it omits the terms of the residual log-likelihood, RL, which are independent of the variance parameters."

The deviance given by `PRINT=deviance` is minus twice the log-likelihood that enters into Genstat's likelihood ratio test statistic. These likelihoods differ for the full and reduced models, and the formulae are as follows:

$$RL_F(\theta) = (y - X_F\alpha_F(\theta))'V_F^{-1}(y - X_F\alpha_F(\theta)) + \log|V_F| + \log|X_F'V_F^{-1}X_F| - \log|X_F'X_F| + (n-p)\log 2\pi$$
$$RL_R(\theta) = (y - X_R\alpha_R(\theta))'V_R^{-1}(y - X_R\alpha_R(\theta)) + \log|V_R| + \log|X_F'V_F^{-1}X_F| - \log|X_F'X_F| + (n-p)\log 2\pi$$

On the other hand, the monitoring deviance is calculated according to the following fomula:

$$-2\lambda(\theta) = (y - X\alpha(\theta))'V^{-1}(y - X\alpha(\theta)) + \log|V| + \log|X'V^{-1}X|.$$

From Genstat output, we have the following information.

full model parameters

$\alpha = 5.734, 0.293, 0.4726, 0.8449, 0.695, 0.8725$

$\theta_1 = 1.421, \theta_2 = 0.157$ i.e. $\gamma_1 = 9.0941$

$\lambda(\theta) = 11.0187$

$RL_F(\theta) = 72.76$

reduced model parameters (maximise adjusted restricted likelihood)

$\alpha = 6.184$

$\theta_1 = 1.488, \theta_2 = 0.298$ i.e. $\gamma_1 = 4.9948$

$RL_R(\theta) = 93.22$

reduced model parameters (hold θ fixed)

$\alpha = 6.181$

$\theta_1 = 1.421, \theta_2 = 0.157$ i.e. $\gamma_1 = 9.0941$

$RL_F(\theta) = 101.4$

test statistics, all χ^2_5 under H_0

$2\log(LR) = 20.45$ p value $= 0.001$

$W = 28.5$ p value < 0.0001

$2\Delta D = 28.5$ p value < 0.0001

9. Summary

Three statistics are available in Genstat 5 Release 3.1 for screening covariates:

$$\begin{aligned} 2\log(LR) &= (y - X_R\hat{\alpha}_R)'\hat{V}_R^{-1}(y - X_R\hat{\alpha}_R) + \log|\hat{V}_R| + \log|X_F'\hat{V}_R^{-1}X_F| \\ &\quad - (y - X_F\hat{\alpha}_F)'\hat{V}_F^{-1}(y - X_F\hat{\alpha}_F) - \log|\hat{V}_F| - \log|X_F'\hat{V}_F^{-1}X_F| \\ W &= \hat{\alpha}_{F1}'[G^{11}_{\alpha\alpha}]^{-1}\hat{\alpha}_{F1} \\ 2\Delta D &= (y - X_R\hat{\alpha}_R)'V_F^{-1}(y - X_R\hat{\alpha}_R) - (y - X_F\hat{\alpha}_F)'V_F^{-1}(y - X_F\hat{\alpha}_F) \end{aligned}$$

All three are asymptotically equivalent, and, assuming normally distributed data, all three have an asymptotic distribution χ^2_s under H_0.

10. References

Genstat 5 Committee (1993) *Genstat 5 Release 3 Reference Manual* Oxford University Press, Oxford.

Harville D A (1974) Bayesian inference for variance components using only error contrasts *Biometrika* **61** 383–385.

Harville D A (1977) Maximum likelihood approaches to variance component estimation and to related problems *Journal of the American Statistical Association* **72** 320–340.

Heritier S and Ronchetti E (1994) Robust bounded-influence test in general parametric models *Journal of the American Statistical Association* **89** 897–904.

Patterson H D and Nabugoomu F (1992) REML and the analysis of a series of crop variety trials *Proceedings of the 16th International Biometric Conference* 77–93.

Richardson A M and Welsh A H (1996) Covariate screening in mixed linear models *Journal of Multivariate Analysis* **58** 27–54.

Welham S J and Thompson R (1997) Likelihood ratio tests for fixed model terms using residual maximum likelihood *J. Royal Statist. Soc. B* **59** 701–714.

The work described in this paper was carried out while Alice was a PhD student in the Department of Statistics at the Australian National University, Canberra. This paper is an informal, fleshed-out version of the slides Alice used in her talk at the Genstat Conference in Wagga Wagga, Australia in December 1994. A more formal version of the same material has now been published as Richardson and Welsh (1996).

Appendix: Genstat code and output

The following Genstat code was used to produce the statistics mentioned in Section 8.

```
open 'wheat.dat'; ch=4
units[nvalues = 46]
factor[levels = 6] variety
factor[levels = 10] centre
variate int; values = !(46(1))
read[ch=4] variety, centre, y
close 4
vcomponents[fixed = variety] random = centre
reml[print = comp, moni,devi, wald; submodel='constant'] y
vdisplay[print = effects]
vkeep[sigma2=ef; rss = rssf] terms = centre; components = bf
calc rssf = rssf/ef
print rssf
vcomponents random = centre; initial = bf, ef; constraints = fixabsolute
reml[print = comp, moni, devi] y
vdisplay[print = effects]
vkeep[rss = rssr]
calc rssr = rssr/ef
print rssr
calc deltad = rssr - rssf
print deltad
stop
```