14th European GenStat Applied Statistics Conference

Programme and Abstracts

GenStat

Agri-Food & Biosciences Institute
(AFBI) Belfast
21st July 2008

Conference Programme

Monday 21st July 2008

0845 - 0915		Registration	
0915 - 0930		Welcome	Alan Gordon & Roger Payne
0930 - 1100		Session 1	Chair Simon Harding
	0930	David Baird	*Microarray analysis in GenStat*
	1000	Steve Langton	*The analysis of complex surveys in GenStat*
	1030	Darren Murray	*GenStat in ecology and the environment*
1100 - 1130		Coffee	
1130 - 1300		Session 2	Chair Alan Gordon
	1130	Graham Hepworth	*Prevalence of salmonella in animal feed mills*
	1200	Phil Brain	*An introduction to microneurography*
	1230	Robert Forrester & Luke Pinner	*Fire severity - a burning issue*
1300 - 1400		Lunch	
1400 - 1530		Session 3	Chair Darren Murray
	1430	Roger Payne	*Hierarchical generalized nonlinear models*
	1500	Jackie Potts & Sarah Wanless	*Fitting generalized linear mixed models with correlated random effects*
	1530	Peter Lane	*Visualization and interpretation of meta-analyses, particularly of rare events*
1530 - 1600		Tea	
1600 - 1730		Session 4	Chair Roger Payne
	1600	Duncan Soutar	*Graphics environments and customization*
	1630	Alex Glaser	*Canonical analysis, biplots and triplots*
	1700	Simon Harding	*Running the NAG Library from GenStat*
During breaks		Poster Session	
		Fiona Boland	*Hierarchical generalised linear models analysis of bovine tuberculosis on milk data*
		Archie K. Murchie & Alan Gordon	*Detecting sex ratio bias in gall midges* (Cecidomyiidae*)*
		Samantha Cook, Elspeth Bartlet, Darren Murray & Ingrid Williams	*The role of pollen odour in the attraction of pollen beetles to oilseed rape flowers*

Advanced Linear Models Workshop Programme

Tuesday 22nd July 2008

Presenters: Roger Payne & Darren Murray

0900 - 1200	Session 1		*Analysis of correlated data:*
			REML recap;
			Spatial analysis;
			Repeated measurements;
			Nonlinear models

1200 - 1300	Session 2		*Design and sample size:*
			power and significance;
			t-tests (1- or 2- sided, inferiority, equivalence);
			analysis of variance

1300 - 1400	Lunch		
1400 - 1700	Session 3		*Counts and proportions with several sources of error variation:* *Generalized linear models recap;* *Generalized linear mixed models;* *Hierarchical generalized linear models;* *Bayesian methods with GenStat-WinBUGS link*

Microarray analysis in GenStat

David Baird
VSN (NZ) Limited,
2 Hardie Place,
RD 2,
Wanaka 9382,
Central Otago,
New Zealend.
Email: David@VSN.CO.NZ

This talk will look at the issues involved in analysing Affymetrix microarrays. These microarrays stretch the limits of the data GenStat can handle, with the largest now containing around 1 million cells on a single chip. The approach used by Affymetrix chips, and how the data is then used and analysed will be outlined. The methods for the Affymetrix MAS 4 & 5 and RMA algorithm's will be demonstrated. The maximum-likelihood solution for the RMA error model will be derived and approaches to compute this quickly in GenStat will demonstrated. The talk will cover techniques generally useful for working with large datasets in GenStat.

Hierarchical generalised linear models analysis of bovine tuberculosis on milk data

Fiona Boland
UCD School of Mathematical Sciences,
University College Dublin,
Belfield, Dublin 4,
Ireland.
Email: fiona.boland@ucd.ie

Hierarchical Generalised Linear Models (HGLM's) are used here to model the relationship between bovine tuberculosis and milk yield data. In broad terms HGLM's extend generalised linear mixed models (GLMM's) by allowing more flexibility in the choice of the distribution for the random effects and also modelling of the dispersion of the random effects and error term.

Mycobacterium bovis (*M. bovis*), the casual organism in bovine tuberculosis (TB) is an important infection of cattle in many countries, especially Ireland and the UK. TB outbreak in a dairy herd causes monetary losses to the exchequer as the infected animals are slaughtered and farmers compensated but it also affects the farmer due to trade restrictions imposed on the whole herd. For the first time in Ireland we examine the effects of bovine TB on milk production and quality. HGLM's are applied to a random sample of Irish dairy herds restricted from trading with at least two or more TB infected animals between the 1st June 2004 and the 31st May 2005. Milk variables (fat, protein, milk yield) belonging to all lactations on an animal in the study are considered and TB infected and non-TB infected animals are compared on these variables. The TB data were obtained from the Department of Agriculture, Fisheries and Food (DAFF) and the milk production data comes from the Irish Cattle Breeding Foundation (ICBF).

There is an inherent hierarchical structure in the data, lactations are nested within animals and animals within herds. Since observations relating to lactations within an animal will be correlated and in addition animals in the same herd will probably be correlated any model will need to incorporate these effects. HGLM's using the estimation method of h-likelihood provide a useful class of models that do this. In addition the variances of the random effects are also modelled. Likelihood criteria were used for both inclusion/exclusion of random and fixed effects to find the best model. Diagnostic plots were examined and the implications of the results of the model for TB were interpreted.

An introduction to microneurography

Phil Brain
Pfizer Global R&D (c096),
Ramsgate Rd,
Sandwich,
Kent,
CT13 9NJ,
UK.
Email: Phil.Brain@pfizer.com

Microneurography studies the voltage responses of electrically-stimulated nerves. The signals can potentially give objective measures of drug effects on the human body, in particular on perceived pain. Measurements are typically made at the rate of 10,000 per second with experiments lasting several hours, so a single experiment on a single patient can give ~ 10^8 measurements. I will give an overview of the subject area and present some initial approaches to analysing the data and detecting meaningful signals. One approach to model these signals involves fitting a sequence of "shark's fins", which demonstrates the power and flexibility of nonlinear regression in Genstat very effectively.

The role of pollen odour in the attraction of pollen beetles to oilseed rape flowers

Samantha Cook[1], Elspeth Bartlet[1], Darren Murray[2] and Ingrid Williams[1]
[1]Rothamsted Research
Harpenden,
Hertfordshire,
AL5 2JQ,
UK
[2]VSN International Ltd,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts,
HP1 1ES,
UK.

The role of pollen odour in resource location by the pollen beetle, *Meligethes aeneus* (F.) (Coleoptera: Nitidulidae), a pollen-feeding insect regarded as a pest of oilseed rape, *Brassica napus* L., (Brassicaceae), was investigated in a linear track olfactometer. Both male and female beetles were attracted to the odour of whole oilseed rape flowers, indicating that these insects can locate their host plants using floral odours as cues. The attractive odour of flowers was found to emanate from all floral parts tested: the petals/sepals, the anthers and from pollen itself. Therefore, at least part of the attractive odour of oilseed rape flowers emanates from pollen. Beetles were more attracted to floral samples containing anthers than those without anthers when these odours were directly compared in a choice-test, and this indicates that there were detectable differences between them. Such differences could be qualitative, and the possibility of pollen-specific odours as pollinator attractants is discussed.

Fire severity - a burning issue

Robert Forrester[1] and Luke Pinner[2]
[1]Statistical Consulting Unit,
ANU, ACT,
Australia
Email: Bob.Forrester@anu.edu.au
[2]Fenner School,
ANU, ACT,
Australia.

Wildfire is a regular problem faced by the forests of south eastern Australia. In 2003 the bushfires in the Kosciusko National Park in NSW were particularly extensive and severe in intensity. The area burnt stretched the full length of the park from the ACT to the Victorian border. Many rare or threatened species of plants and animals were affected.

Fire severity was assessed for the entire area that was burnt and the aim of the study was to model this severity using a range of explanatory factors and variates. Among the possible explanatory variates of interest were vegetation type, fuel load, elevation, slope, time since last fire and the ffdi (forest fire danger index) at 9:00am and 3:00pm.

Since the data were available for the entire national park at the pixel level (25m x 25m) this posed a number of problems. Amongst these were the sheer quantity of data (about 28 million items) and the spatial and temporal correlation present. The analysis proposed fits a proportional odds model to a subset of the data in order to explore the main explanatory variates affecting the fire severity.

Reference

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models (second edition)*. Chapman & Hall, London.

Canonical analysis, biplots and triplots

Alex Glaser
VSN International,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts,
HP1 1ES,
UK.
Email: alex.glaser@vsni.co.uk

Canonical analysis is the simultaneous analysis of two, or more, data tables. This is particularly relevant to ecological analysis where we often have tables of species levels and environmental descriptors observed at the same locations. Two techniques, redundancy analysis (RDA) and canonical correspondence analysis (CCA), have been added in the 10th and 11th Editions of GenStat *for Windows*. These techniques allow you to study relationships between the objects, numerically and graphically: i.e. in a species/site table they allow you to establish which species and sites share common traits with different environments.

Examples from Legendre and Legendre (1988) and the RDA and CCA procedures will be used to show how to perform canonical analysis, and produce biplots and triplots of environmental, site and species scores to aid interpretation. We also show how known effects can also be "partialled" out, or equivalently, how environmental descriptors can be isolated to show their individual effects.

Reference

Legendre, P. & Legendre, L. (1998). *Numerical Ecology, Second English Edition*. Elsevier, Amsterdam.

Running the NAG Library from GenStat

Simon Harding
VSN International,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts,
HP1 1ES,
UK.
Email: simon.harding@vsni.co.uk

The 11th Edition of GenStat *for Windows* contains a new directive NAG that allows you to run over 80 algorithms from the Numerical Algorithms Group’s Library. These include algorithms for finding the zeros of polynomials, calculating integrals, solving ordinary differential equations and solving linear and quadratic programming problems. This talk will explain how to specify the input and output arguments for the NAG algorithms within GenStat, and will use some of the many examples accessible from the new Examples menu to show how this link allows GenStat now to be used to solve mathematical as well as statistical problems.

Prevalence of salmonella in animal feed mills

Graham Hepworth
Statistical Consulting Centre,
The University of Melbourne,
Australia.
Email: g.hepworth@ms.unimelb.edu.au

Salmonella bacteria can cause different types of illness in humans, the most common being gastroenteritis. Most salmonella outbreaks are caused by contaminated food of animal origin. Animal feed mills have been recognized as a potential source of infection, and so the stock feed industry throughout the world has seen the importance of implementing appropriate control measures.

In 2002 a set of procedures and standards known as the FeedSafe program was introduced to Australian stock feed manufacturers. Similar programs have appeared to work well in other industries, but there was limited data available to assess its effectiveness with animal feed. A survey was conducted between 2003 and 2007 on 17 feed mills, with a total of 4500 samples submitted for laboratory analysis. Samples were taken from raw materials, finished feeds and milling equipment. The main questions of interest were:

- Is the prevalence of salmonella changing over time?
- Does time of year influence prevalence?
- Where are the contamination points in the feed mills?
- Is salmonella prevalence influenced by site of sample (e.g.pellet press, cooler), type of sample (e.g. finished feed, meat meal) or geographical location?

The results were mostly encouraging for the industry, with a few challenges. The main tool of analysis was the fitting of generalized linear mixed models in GenStat, with the main issues being how to categorize the explanatory variables and how to specify the random effects.

Visualization and interpretation of meta-analyses, particularly of rare events

Peter Lane
Research Statistics Unit,
GlaxoSmithKline,
Harlow,
UK.
Email: Peter.W.Lane@gsk.com

Meta-analyses are increasingly being used to summarize information across clinical trials, often to publicize good or bad news. Public access to trial results on the Internet has made it especially easy to generate such meta-analyses, particularly of safety issues. Once the hurdles of acquiring and selecting data have been cleared, the task of analysis with some given technique is only too easy, but the results are often poorly presented. I will describe a range of graphs designed to summarize and help interpret meta-analysis, all of which are straighforward to produce in GenStat (though the circular axis of the Galbraith plot is a little challenging). One of the key issues is the choice of scale for the analysis (such as the log-odds scale), which should be driven by the underlying science, and the often different scale (such as the risk scale) on which the results need to be interpreted. Rather than attempting to re-analyse on the interpretation scale, as is often suggested, the natural approach is to form predictions from the scientifically reasonable model. I will look specifically at the fixed-effects meta-analysis of a binary response, illustrated by publicly available data from last year's high-profile analysis of Avandia with respect to cardiovascular safety.

The analysis of complex surveys

Steve Langton
Dept for Environment, Food & Rural Affairs / Steve Langton (Statistical Consultant)
Email: stats@slangton.org.uk

Defra's farm survey team conducts a variety of surveys of agricultural holdings in England, mainly using stratified random surveys and ratio estimation (Sampford, 1962). Until 2001, the surveys were analysed by a Microsoft Access program, which was difficult to use and produced only basic output. Various packages, including general statistics software and survey analysis programs, were considered as alternatives, but none provided exactly what was needed, and so a series of GenStat procedures were written to manipulate the data and produce the required ratio analysis.

The procedure SVSTRATIFIED used for the ratio analysis was of general applicability and was submitted for inclusion in the procedure library of the 8th Edition of GenStat *for Windows*. Besides producing estimates of means, totals and ratios, SVSTRATIFIED produces graphs and other output to allow identification of outliers. SVSTRATIFIED displays influence statistics defined as the percentage change in the overall estimate when an observation is replaced by a missing value. The influence statistics can also be used as a criterion for prioritising investigation of anomalous observations (see for example, Lawrence & McKenzie 2000). A feature of SVSTRATIFIED is that the data may be supplied either as one value for each sampled (or responding) unit, as in most other survey packages, or as one value for each unit in the population, with missing values for unsampled units. In situations where a complete database of the survey population is available, the latter format makes specifying the design simpler, since information such as the number of units in each stratum does not have to be explicitly declared.

In the 9th Edition a number of other survey procedures were introduced in order to make GenStat more generally useful to survey statisticians. The most important of these was SVTABULATE, which is designed to produce cross-tabulations, as with TABULATE, but with the correct standard errors allowing for the design of both one- and two-stage sample surveys. For standard designs SVTABULATE does not need weights to be explicitly provided, but a number of procedures are available to manipulate them in more complex applications. SVWEIGHT creates weights for standard designs, SVREWEIGHT modifies them to adjust for outliers, whilst SVCALIBRATE allows calibration of weights (Deville & Sarndal 1992).

Since then a number of further improvements have been made. Both SVSTRATIFIED and SVTABULATE now allow bootstrapping as an alternative to the usual Taylor series variance estimators. Influence statistics are produced for all estimated means, totals or ratios, rather than just for the grand total. Percentiles can be estimated with bootstrapped standard errors. SVGLM fits generalised linear models to survey data; bootstrapping is used for non-normal models and for prediction, whilst in the normal case the Taylor-series approximation is also available.

Further enhancements are planned, including a survey analysis guide. The views of users on the features they require are useful in prioritising new developments.

References

Deville, J.-C. & Sarndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Lawrence, D. and R. McKenzie (2000), The General Application of Significance Editing, *Journal of Official Statistics*, **16**, 243-253.

Sampford, M.R. (1962). *An introduction to Sampling Theory*. Oliver and Boyd, London.

Detecting sex ratio bias in gall midges (*Cecidomyiidae*)

Archie K. Murchie and Alan Gordon
Applied Plant Science & Biometrics Division,
Agri-Food & Biosciences Institute,
Newforge Lane,
Belfast,
BT9 5PX,
UK.
E-mail: archie.murchie@afbini.gov.uk

Gall midges (*Cecidomyiidae*) are small plant-feeding flies that are very common but often overlooked unless they are crop pests. When rearing through some cecidomyiid species, entomologists have noted that the sex of broods from individual females can be highly biased to one sex. In some cases, females produce unisexual broods (all of one sex). The mechanism by which this happens involves elimination of paternally derived sex chromosomes. Whether or not this is mediated by bacterial intracellular parasites (e.g. *Wolbachia*) is not clear. When broods are large and the sex bias strong then monogeny (production of sex-biased broods) is clearcut. However, when flies emerging from broods are few (<5) and brood sex mixed, detecting monogeny is more awkward. We are currently looking at methods to detect evidence of monogeny in gall midge species, using randomisation tests written in GenStat.

GenStat in ecology and the environment

Darren Murray
VSN International,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts,
HP1 1ES,
UK.
Email: darren.murray@vsni.co.uk

GenStat contains many tools for measuring ecological and environmental data. In this talk facilities for measuring biodiversity such as species abundance distributions, and diversity and evenness statistics are introduced. An important part of species diversity is the number of species in a community. GenStat contains methods for estimating the species richness using non-parametric methods, and has facilities for generating and modelling species accumulation or rarefaction curves. Another feature of ecological data is that many measurements are recorded as counts. Sometimes the number of zeros recorded can exceed what would typically be expected if modelling using a Poisson or negative binomial model. A new procedure will be included in the next release for fitting zero inflated Poisson and negative binomial models to count data with excess zeros.

Hierarchical generalized nonlinear models

Roger Payne
VSN International,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts, HP1 1ES, UK.
Email: roger.payne@vsni.co.uk

Hierarchical generalized linear models (HGLMs) provide a convenient and efficient way of analysing counts and proportions when there are several sources of error variation. So, for example, they can be used to analyse counts and proportions that may be observed from split-plot experiments, or from medical trials involving several centres or patient groupings. They extend the familiar generalized linear models (GLMs) by allowing you to include additional random terms in the linear predictor. However, they do not constrain these terms to follow a Normal distribution nor to have an identity link, as is the case in the more usual generalized linear mixed model (GLMM). So they provide a richer of class of models that may be more intuitively appealing. The HGLM algorithm involves fitting two (or more) interlinked generalized linear models, firstly to estimate the fixed and random effects in the model that describes the mean, and secondly to model the dispersion of the random terms. So all the familiar model checking techniques are available. See Lee, Nelder & Pawitan (2006) or Section 3.5.11 of Payne *et al.* (2008) for further details.

Another extension to GLMs allows for the inclusion of nonlinear parameters in the linear predictor (see Lane 1996, or Section 3.5.8 of Payne *et al.* 2006). The fitting algorithm for these generalized nonlinear models operates by performing a nested optimization, in which a generalized linear model is fitted for each evaluation in a optimization over the nonlinear parameters. The optimization search thus operates only over the (usually relatively few) nonlinear parameters, and this should be much more efficient than a global optimization over the whole parameter space. This talk will explain how similar principles can be used to include nonlinear fixed parameters in the mean model of an HGLM, thus defining a *hierarchical generalized nonlinear model*. The methods will be illustrated using the analysis of data from Stuart Cooney of ANU studying the relationship between Hooded Parrot (*Psephotus dissimilis*) nestlings and moth larvae.

References

Lane P. (1996) Generalized nonlinear models. In: *COMPSTAT 1996 Proceedings in Computational Statistics*, pp. 331-336.

Lee, Y., Nelder, J.A. & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. CRC Press.

Payne, R.W., Harding, S.A., Murray, D.A., Soutar, D.M., Baird, D.B., Welham, S.J., Kane, A.F., Gilmour, A.R., Thompson, R., Webster, R. & Tunnicliffe Wilson, G. (2008). *The Guide to GenStat Release 11, Part 2 Statistics*. VSN International.

Fitting generalized linear mixed models with correlated random effects

Jackie Potts[1] and Sarah Wanless[2]
[2]Biomathematics and Statistics Scotland,
The Macaulay Institute,
Craigiebuckler,
Aberdeen,
AB15 8QH,
UK
Email: j.potts@macaulay.ac.uk
[2]Centre for Ecology and Hydrology
Bush Estate,
Penicuik,
Midlothian,
EH26 0QB,
UK.

Recent modifications to the Genstat GLMM procedure allow the fitting of Generalized Linear Mixed Models with correlated random effects, although a limitation is that models with correlated errors cannot be fitted. An example is presented in which a random coefficient regression model is fitted to binary data on the breeding success of seabirds. The slope of the regression on the number of previous breeding attempts by each pair is allowed to vary between pairs.

A simulation study is presented to investigate how well the parameters can be estimated in GLMMs with a binary response and correlated random effects.

Graphics customization and enhancements

Duncan Soutar
VSN International,
5 The Waterhouse,
Waterhouse Street,
Hemel Hempstead,
Herts,
HP1 1ES,
UK.
Email: duncan.soutar@vsni.co.uk

Graphical environments are a new feature in the 11th Edition of GenStat *for Windows* that give users a convenient way to alter the default appearance of graphs. Another new feature is the ability to display user-supplied information when using the "Data Information" tool in the graphics viewer. These enhancements will be demonstrated together with other new features, including control over the order in which data sets are plotted, scatter-plots with linked axes and a method for displaying 2-dimensional bitmaps.