CASI-1991

Proceedings Booklet

Invited Speakers

Prof. John Haslett, TCD, Ireland
Prof. David Clayton, MRC Biostatistics Unit, UK
Prof. Sir David Cox, Nuffield, Oxford, UK
Prof. Peter Green, Bristol, UK.

SPATIAL DATA ANALYSIS - CHALLENGES

John Haslett

Department of Statistics
Trinity College Dublin

A number of studies involving spatial data are reviewed and an attempt is made to extract some common themes. We discuss both exploratory data analysis and modelling.

Particular emphasis will be given to studies in which some measured phenomenon, such as wind speed or geochemistry, varies over space (and perhaps in time). In this context there is always a need to explore the data and to describe the variation. Often it is sufficient to give this description in terms of patterns and exceptions. Sometimes there is an additional need to express these in formal terms with the aid of a model. As with all modelling there is a consequent need to be aware of lack of fit and related ideas. In broad terms this is an approach that is used in many areas of data analysis. What are the special features of spatial data analysis?

There are considerable parallels with time series data. Frequently the data are observational. Often stationary concepts are inadequate. Usually there is context. Typically in space the data are irregularly distributed. There may well be data of a variety of types pertaining to the same region. These concepts will be illustrated.

What modern tools are available? Considerable advances in computer graphics have revolutionised the exploratory approach. The most widely used spatial modelling procedures are those based on geostatics or close cousins. Diagnostic analysis of the lack-of-fit of such models is however in its infancy. It will be seen that there are many unresolved issues.

A MATHEMATICAL MODEL OF THE LACTATION CURVE OF THE DAIRY COW TO INCORPORATE METABOLIZABLE ENERGY INTAKE

S D Lennox[1], E A Goodall[1] and C S Mayne[2]

[1]Department of Biometrics
The Queen's University of Belfast, Newforge Lane
Malone Road, Belfast BT9 5PX

[2]Agricultural Research Institute of Northern Ireland
Hillsborough, Co Down, BT26 6DR

This paper models the underlying lactation curve of the dairy cow in terms of total energy intake. Algebraic models were initially fitted to data from the Agricultural Research Institute of Northern Ireland to examine the effect of diet on milk yields. Mean values of the parameter estimates of the models were then regressed against corresponding values for energy intake. The resulting relationships were incorporated in a new model defining winter milk yields in terms of energy intake.

A NEW APPROACH TO AHP DECISION-MAKING

H A Donegan[1], F J Dodd[1] and T B M McMaster[2]

[1] Department of Mathematics,
University of Ulster

[2] Department of Pure Mathematics,
The Queen's University of Belfast

Saaty's Analytic Hierarchical Process is one of the most commonly used methods of prioritising the elemental issues in a complex problem. This paper goes some way towards overcoming some objections to the method and offers an explanation as to why the method is more suitable for simple prioritisation rather than for the much more useful task of the quantification of weightings for the issues. It suggests a strategy for overcoming this limitation and, incidentally, indicates how the application of the method might be extended to cater for the case of judgements determined by consensus among a panel rather than by a single individual.

ROBUST REGRESSION ESTIMATORS - THE CHOICE OF TUNING CONSTANTS

Gabrielle E Kelly

Department of Statistics
University College Dublin
Belfield, Dublin 4

The robust regression estimators of Huber and Welsch and the bounded influence estimators of Krasker and Welsch require the specification of a cut-off or tuning constant before they are fully defined. Here the asymptotic mean squared errors of these estimators under different designs/distributions for the explanatory variables and different error distributions is computed. The choice of tuning constants is seen to be critical in the trade-off between bias and variance. The choice illuminates the differences in behaviour of the estimators. Two numerical examples further illustrate the differences.

SOME RECENT DEVELOPMENTS IN ECONOMETRIC MODEL SELECTION AND ESTIMATION

[1]Derek Bond and [2]Michael Harrison

[1]Ulster Business School, University of Ulster at Jordanstown
Shore Road, Newtownabbey, BT37 OQB

[2]Department of Economics, Trinity College Dublin

Not Provided

GENERALISED LINEAR MIXED MODELS IN BIOSTATISTICS

D Clayton

MRC Biostatistics Unit, Cambridge

Generalised Linear Mixed Models (GLMMs) are generalised linear models which include one or more random effect. They are extremely important in the analysis of data concerning repeated measures on the same subjects, and there has recently been considerable progress in developing estimation procedures for such problems. Interestingly, problems which involve fitting non-parametric "smooth" relationships share a common mathematical structure.

The paper will briefly describe three approaches to inference in this very general class of models, and review applications in biostatistics.

IMPROVEMENTS IN INTEGRAL EQUATION MODELS FOR ESTIMATES OF THE LEVEL OF HIV INFECTION IN IRELAND

C M Comiskey

School of Mathematical Sciences, Dublin City University

Linear Volterra integral equations of the first kind have been used in the past to provide estimates of the level of HIV infection in Ireland. Such models can easily be solved in some circumstances .

However, under other circumstances these models prove to be more difficult to solve for *h(t)*, the numbers of HIV cases in year *t*. Although the basic equation may be viewed as the convolution of h and Ÿ we cannot evaluate it by using inverse Laplace transforms. Nor can we transform to form a Volterra equation of the second kind.

We show how the integral equation model can be changed to a generalised Abel integral equation, the solution of which is given in terms of a new integral. We subsequently evaluate this integral in terms of a Gamma function plus a remainder in the form of a series in t or an incomplete Gamma function. We also provide error bounds for the remainder. This new solution allows us to predict new and more reliable estimates of the level of HIV infection in the Irish population.

ETS AND LUNG CANCER: A CRITIQUE OF A META-ANALYSIS

S J Kilpatrick

Medical College of Virginia,

Virginia Commonwealth University

Wald et al, BMJ 293: 1217, 1986 conclude from a meta-analysis of 10 case control studies and three prospective studies that there is a significantly high odds ratio between spousal smoking and lung cancer in nonsmokers.

This paper shows that an adjustment for country of origin negates this conclusion as does the addition of one study not included by the original authors.

THE APPLICATION OF THE MODIFIED COVARIANCE TECHNIQUE IN THE ANALYSIS OF DATA RELATING TO TREATMENT OF CANCER OF THE BLADDER

[1]E S Gillespie, [2]L Stewart and [3]R N Wilson

[1] Department of Mathematics, University of Ulster, Jordanstown
[2] Department of Urology, Belfast City Hospital
[3] Computer Services Department, University of Ulster, Jordanstown

Serum levels of the nucleotide salvage enzyme, thymidine kinase (TK), have been assayed in patients with various stages of cancer of the bladder and patients without cancer.

The prime aim is to determine whether or not the patients have significantly higher levels of TK than the control patients as has been found in patients suffering from other forms of cancer (1,2).

To date, for cancer of the bladder, the number of patients and in some cases, the number of controls have been small. Also, there are errors involved in the measurements of differential white cell counts of the peripheral blood and homonuclear leucocytes which must be taken into account.

To enhance the quality of the control data, X, the prediction matrix

$P = X(X'X)^{-1}X$

has been applied to detect those controls whose n measurements indicate that they are far removed from the rest of the control group. Allowances have been made for errors in measurement through the n diagonal elements of the covariance matrix, ($X'X$).

An iterative procedure has been devised to enhance the quality of the control group.

Key References

(1) McKenna P G et al, Br J Cancer 1988; 57: 619-622.

(2) McKenna P G et al, J Clin Hematol Oncol 1985; 15: 71.

A TEST OF SEN'S ENTITLEMENT HYPOTHESIS

P McGregor, I Cantley

University of Ulster, Jordanstown

Not Provided

THE ROLE OF STATISTICAL MODELS

D R Cox

Nuffield College, Oxford

Most relatively advanced use of statistical methods involves a probability model representing a process producing the data. Various kinds of such model are distinguished in part via their depth of contact with the underlying subject matter. In particular, a primary distinction is drawn between substantive models and empirical models. Examples are given from various fields of study ranging from the physical sciences through epidemiology to the social sciences. The implications for applied statistical work are stressed.

A MODEL FOR BIVARIATE FAILURE TIMES SUBJECT TO RIGHT CENSORING

G MacKenzie

Department of Epidemiology and Public Health,
The Queen's University Belfast

The statistical analysis of failure-time data has been advanced by developments in the study of bivariate models. Following Cox's (1972) seminal paper on the proportional hazards model, Clayton (1978) proposed a bivariate model designed to explore the association between the age of death of fathers and their sons. Oakes (1982) clarified some issues relating to Clayton's partial likelihood. More recently, Hougaard (1984) derived a new bivariate failure time model by introducing unobservable frailty distributions. This device was employed by Clayton and Cuzick (1985) to obtain a bivariate generalisation of the original proportional hazard model given by Cox.

Distributional assumptions of one form or another are required in the derivation of each of the models noted above. Relatively speaking, these derivations are mathematically non-trivial, and their generalisation to more than two dimensions does not appear straightforward. Such problems raise questions about their accessibility and widespread acceptance.

Accordingly, it would seem sensible to consider alternative methods based on potentially less restrictive assumptions. The purpose of the present paper is to outline a distribution-free procedure for the analysis of bivariate survival times.

Let T_1 and T_2 denote the pair of failure-time random variables involved. The method is based on a partition of the total period of observation (O,T) into k intervals (t_{i-1},t_i) for $i=1,...,k$ and where $t_o = 0$ and $t_k = T$. Let the unconditional probability that T_1 fails during the i-th. interval and T_2 fails during the j-th. interval be:

$$p_{ij} = pr(T_1 = i, T_2 = j) \qquad (1)$$

It is shown that the likelihood can be written in terms of (1) in the uncensored and censored case subject to $\Sigma\Sigma p_{ij} =1$. The form of the likelihood under the independence assumption is obtained and the methods are applied to example data.

Key References

Clayton D G (1978). A model for association in bivariate life tables and its application in epidemiologic studies of familial tendency in chronic disease incidence. Biometrika 65: 141-151.

Clayton D G, Cuzick J (1985). Multivariate generalisations of the proportional hazards model. JRSS A 148: 82-117.

INDEPENDENT TRIALS ARE A MODEL FOR DISASTER

D A Jackson

Department of Statistics
Trinity College Dublin

Introduction

Systems fail. Where a system has a number of back-up systems, a failure is usually no more than a temporary inconvenience. A disaster however may result from a series of failures of these back-up systems. Where reliance is made on a series of back-up systems in order to lessen the probability of disaster, assumptions regarding the independence of these systems may be made, leading to solutions stating numbers of back-up systems required to bring the probability of disaster below any given level.

For a system with a series of back-up's all of which have to fail in turn for a disaster to occur, we show that the existence of a simple type of dependence between the failure rates can lead to vast increases in the probability of disaster relative to the independence model.

Model

Denote the ith back-up by Xi. Failure of Xi increases the odds for a failure of Xi+1 by a constant factor k.

Let P_i = Prob. that *i*th back-up fails, given that the first *i*-1 fail.
Let O_i = Odds that *i*th back-up fails, given that the first *i*-1 fail.

By definition $O_i = P_i/(1 - P_i)$ (1)

Then $O_i = kO_{i-1}$ (2)

The type of dependence we are postulating is shown to exist in nature in for example contests that are decided, not by a single trial, but by a series of trials (supposedly identical). In many sports, for example Tennis, heavy defeats, or disasters are observed to be more common than an independent trials model would predict. A model which increases the odds by a constant factor for failure in the next trial, if one has lost the previous trial, provides excellent fits to this type of data.

PREDICTING THE GROWTH OF MANPOWER FOR THE NORTHERN IRELAND SOFTWARE INDUSTRY

S McClean, P Lundy, W Ewart

University of Ulster

Not Provided

STOCHASTIC ALGORITHMS AND BAYESIAN INFERENCE

P J Green

Department of Mathematics
University of Bristol

Stimulated mainly by the requirements of statistical model-based approaches to image analysis, there have been recent developments in simulation procedures appropriate for models containing many random variables interacting in a complex fashion. In the image analysis context, such stochastic algorithms are used both for image synthesis: constructing realisations from stochastic image models, and image analysis: making Bayesian inference about a true image on the basis of an observed degraded version. In contrast to deterministic methods, stochastic algorithms are easily adapted to perform inference more subtle than simple image restoration, including interval estimates and tests for anomalies.

These methods have potential for other areas of statistical methodology where dependence is crucial, including the classical problem of calculating marginal posterior distributions in multiparameter Bayesian inference.

HOW INTERACTIVE GRAPHICS WILL CHANGE STATISTICAL PRACTICE

A Unwin

Department of Statistics
Trinity College Dublin

Interactive graphics have changed how statisticians explore data. They will also change how statisticians work and communicate with others. Through live visual displays interactive graphics offer insights which are impossible to achieve with traditional statistical tools. Statisticians can more understandably explain what they do, while domain experts can more readily contribute their subject knowledge to data analyses.

HIGH-INTERACTION DIAGNOSTICS FOR GEOSTATISTICAL MODELS

OF SPATIALLY REFERENCED DATA

R Bradley, J Haslett

Department of Statistics,
Trinity College Dublin

The aim of post-modelling diagnostics has traditionally been to perform a model criticism. In addition to this, those presented in this paper also allow the user to explore the data using the model, to decompose the model spatially and to detect anomalies. Development of diagnostic tools for spatially referenced data, a type of data that often arises in the geological sciences, has been greatly neglected. This is because until recently there has been no effective way of displaying and analysing the plots of such diagnostics. The high interaction graphics have been used to great effect in general exploratory data analysis (Cleveland and McGill, 1988). It has recently been adapted for use with spatially referenced data (Haslett et al, 1991).

This paper extends the use of the paradigm to the area of post-modelling diagnostics. Interactive graphics give a platform to display and manipulate large quantities of statistical information. This can include pointwise, pairwise and subsetwise statistics. With these resources in mind, a theoretical study of diagnostics for spatially referenced data is presented. This involves examining, decomposing and perturbing the model variance/ covariance matrix. Some novel statistics are presented and these and traditional diagnostics are compared using generated data with anomalous effects added.

Key References

Cleveland W S, McGill M E (1988) Dynamic Graphics for Statistics, Wadsworth/
Brooks Cole, Monterey, CA

Haslett J, Bradley R, Craig P, Unwin A, Wills G (1991), 'Dynamic Graphics for Exploring Spatial Data', American Statistician, to appear August 1991.

NOVEL GRAPHICAL METHODS FOR EXPLORATION OF SPATIAL DATA WITH TIME INFORMATION

G J Wills, A R Unwin

Department of Statistics,
Trinity College Dublin

There are several types of spatial data with either explicit or implicit time information; repeated observations of a spatial/temporal point process provide an example of the former and network flow data provides an example of the latter. The analysis of such data is difficult via simple hypothesis testing techniques or by examining global statistics. The authors therefore suggest several high-interactive tools for exploratory analysis of such data. The concept of linked plots, in which the utility of simple univariate or multivariate plots is increased by allowing selection of one plot area to be reflected in all plots, is generalised to allow linking between different types of data; in this case between the network nodes and the associated connecting lines. Based on nearest-neighbourhood methods proposed for use on static point processes, the p-Surface tool for exploratory analysis of moving point processes is introduced and its implementation described. Some examples of spatial-temporal data are shown, and the application of the above techniques to these examples is presented.

PARAMETER ESTIMATION VERSUS CURVE FITTING

'NEW LAMPS FOR OLD'

J J McKeown, D Sprevak

Department of Engineering Mathematics
Queen's University Belfast

We discuss in this paper a methodology for assessing the quality of solutions for the estimates of nonlinear models. The ideas are illustrated by reference to a problem that has been discussed at length in the literature.

SEARCH FOR EFFICIENT LOCAL KRIGING ALGORITHM

M Dillon

Department of Statistics,
Trinity College Dublin

Interpolation is frequently performed by kriging. There are both global and local versions of the procedure. The global algorithms have the advantage of having to calculate the kriging matrix only once. However, a pre-computation time of O(n3) is not efficient. The local variations use the global algorithms on a smaller data set (typically those data points within some radius around the interpolation site). This reduces the total time of the algorithm without much loss of accuracy in the interpolated value.

Local algorithms require two steps at an interpolation site. Firstly, the search for the neighbourhood of the interpolation site. Secondly, the application of the kriging algorithm to find the weights for the interpolation function. In all practical problems these steps will have to be applied repeatedly; for each site to be interpolated.

This paper is concerned with outlining methods to efficiently update both the interpolation neighbourhood and the kriging weights when moving from one interpolation site to another. Preliminary results are given for the determination of the local neighbourhood. The results compare a Voronoi tessellation and a lexographical ordering of the data points. Both structures are assessed using both random and gridded interpolation sites. Iterative solutions for the updating of the kriging weights are also discussed.