Contents

Statistical and Health Issues Arising from BSE and nvCJD

Sheila Gore[1]

[1] MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge, CB2 2SR

Abstract

After a brief historical account which includes statistical aspects of quality control at abattoirs, three current statistical issues of public health importance are addressed;

1. maternal transmission from BSE-dam to calf;
2. whether age-related consumption of meat products is sufficient to explain the younger age of nvCJD cases; and
3. design considerations in estimating prevalence of pre-clinical nvCJD.

Reduction In The Number Of Control Varieties In Distinctness Testing

Sally Watson[1] and Colin Weatherup[2]

[1] Biometrics Division, Department of Agriculture for N. Ireland and Biometrics Department, Queens University, Belfast and [2] The Open University

Abstract

Before a new variety of an agricultural crop can be accepted for general use it must be shown to be distinct (D) from all existing varieties of the crop, it must be uniform (U) in that its plants must be reasonably similar and it must be stable (S) from generation to generation. To enable these assessments to be made spaced plant trials are conducted in which both the candidate and all existing varieties, the controls, are grown in a randomised block design and a large number of characteristics are measured on the individual plants during the growing season. DUS is generally determined from 3 years of recordings with different plants in each year. The talk will mainly deal with the distinctness testing of the ryegrass species for which the Plant Testing Station, Crossnacreevy, is the UK test centre. In these trials each variety is represented by 60 spaced plants on which some 20 characteristics are measure and there are currently some 1150 varieties grown. However as this is a continuing process and each accepted candidate variety becomes a control for later candidates, the number of varieties which must be accommodated in the trials is ever increasing.

The problem of organising the trials to avoid exceeding reasonable limits of size is discussed and an approach which enables candidate varieties to be compared with all other varieties in a reduced number of plots, without any appreciable effect on the stringency of the distinctness test, is described. This requires that each control variety is omitted from tests in one year out of three in a cyclical basis, compensating for this loss of data by using data from earlier years.

With regard to the method of analysis the usual linear model for n_v varieties and n_y years

$$c_{ij} = \mu + y_j + v_i + \varepsilon_{ij} \quad \text{.....(1)}$$

where c_{ij} is the value on a character for variety I in year $j, i = 1, \ldots, n_v$ and $j = 1, \ldots, n_y$

v_i is the effect of the ith variety with $\sum v_i = 0$
y_j is the effect of the jth year with $\sum y_j = 0$
ε_{ij} is a random error associated with variety i in year j
is found to be inadequate for many of the measured characters due to a strong dependency on the 'earliness' of the years. Instead the non-linear modified joint regression analysis (MJRA) model

$$c_{ij} = \mu + y_j + \beta_j v_i + \varepsilon_{ij} \quad \text{.....(2)}$$

is used where β_j is the slope of variety means in year j against variety means over all years and the other symbols are as before. This model was originally proposed by Digby, P (1979) to allow for varying slopes of the means of one variety versus means over all varieties and has been adapted to allow for varying slopes of variety means in one year versus means over all years. It is found that the stringency of the new system with compensated data is similar to the original system with complete data.

References

Digby, P (1979). Modified Joint Regression For Incomplete Variety x Environment Data. *Journal of Agricultural Science 93* Cambridge, 81-86.

Some Designs for Mixtures of Several Plant Species

John Connolly[1], Dieter Ramseier[2] and Fakhri Bazzaz[3]

[1] Department of Statistics, NUID University College Dublin, Dublin 4, Ireland, [2] Geobotanisches Institute ETH, Gladbachstrasse 114, CH-8044 Zuerich, Switzerland and [3] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA, 02138, USA

Abstract

Experiments involving more than two plant species are rare despite the frequent occurrence of multi-species assemblages in nature. This is perhaps not surprising given the logical and statistical problems in effectively dealing with even two species. Various designs have been proposed for two species experiments including replacement (a monoculture of each species at a common density and a mixture with each species at half this density) and additive series, addition series (replacement series at several densities) and factorial designs (with treatment factors the densities of both species but excluding the (0,0) combination). The former two designs have logical difficulties while the latter two designs allow fitting response models for each species which potentially avoid those difficulties. A problem with the addition series and factorial designs is the rapid increase in design size as the number of species increases. Also design size is increased by the inclusion of monocultures for which responses are available only for the species in the monoculture and so are only of use for the response model for that species. For many questions of interest in competition investigations monocultures are not essential and it is to such situations that this work is addressed.

For investigations based on multispecies communities three Types of Simplex design are considered, the Simplex Lattice $\{q, m\}$, the Simplex Centroid and Axial designs. Designs all of whose points are interior (the mixture includes all species) are particularly suitable for plant competition experiments since each mixture contains each species and so there will be a response available for all species from all mixtures. A second consideration is that the design should cover the area of interest well. Most of the Simplex Lattice $\{q, m\}$, the Simplex Centroid designs considered had a low percentage of design points containing all species and also gave a relatively poor cover of the design space. For these reasons Axial designs are often

to be preferred in competition experiments where the separate species responses are measured, particularly if all design points are interior (contain all species). In addition, 0 designs based solely on proportional inclusion of species are subject to various logical difficulties so there must be some repetition at a number of overall densities.

Mixture amount axial designs. The class of design proposed is an axial design with all points internal and repeated at two overall densities T and $T(1+k)$ (in terms of seedling density or initial biomass or leaf area etc.). For q species the simplex has $2q+1$ mixtures (for $q \leq 2$) (see Figure 1 for an example with q = 3). Firstly, there are q mixtures, each dominated by one species with initial proportion β and the each other species at proportion $(1-b)/(q-1)$. Another q mixtures are included, each of which

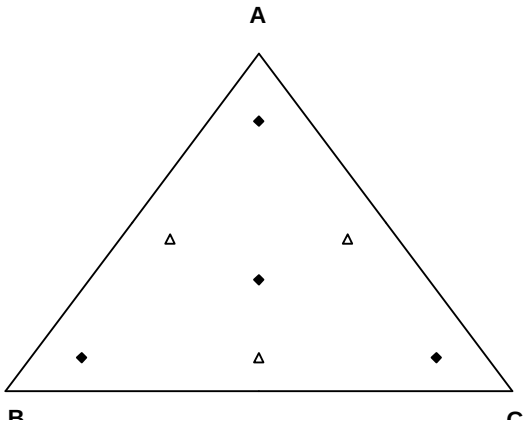

FIGURE 1. Simplex design for 3 species

contain one species at a proportion $(1-\beta)/(q-1)$ and the others at a proportion $(\beta+q-2)/(q-1)2$. Each of these mixtures is the average of q-1 of the initial q mixtures. The centroid with a proportion 1/q of each species is also included. There is an endless variety of possible designs that could be evaluated but here attention is confined to those characterised by T, k and β.

The class of designs examined has the advantages that design size increases only linearly with number of species, the designs cover much of the area of interest for most of the questions addressed by such experiments, all design points contain all species and hence contribute to the estimation of response models for each species and they are easy to use, depending on three readily understandable parameters.

Efficiency of mixture amount designs of the type proposed. Assume that the simple response model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q + \varepsilon$$

holds and that the variance of y is constant at all design points. How does the placement of the design points affect the efficiency with which the design estimates the parameters of this model. Two factors will have a direct bearing on the efficiency, the closeness of the design points to the vertices of the simplex and the distance apart that the densities are set. This design is completely symmetric in all species and so the estimate of the response to the density of any species has the same variance as for any other species, as is also true for the covariance between any pair of estimates.

A general form of the calculation of the efficiency of this design for any given values of β and k is derived. The effect of varying β and k on the standard error of difference between regression coefficients is shown for varying numbers of component species.

Additional statistical difficulties include analysis of correlated responses, the responses of the different species in a mixture, and design and analysis of nonlinear and heteroscedastic responses. There are also problems with optimal design for the model above if the number of design points is to be reduced and of defining models which include nonlinear polynomial terms.

The use of sampling theory to estimate discarding from fishing vessels

Michelle Allen and David Kilpatrick[1]

[1] Biometrics Division, Department of Agriculture for Northern Ireland, Newforge Lane, Belfast BT9 5PX. E-mail: Michelle.Allen@dani.gov.uk

1 Introduction

Whenever a fishing vessel makes a haul the total catch can be separated into commercially desirable fish and discarded fish. The commercially desirable fish are sold but discarded fish are thrown back into the sea. It is an EC statutory requirement that fishing vessels provide data about the quantities of fish landed but not fish discarded. The estimation of discarding practices of commercial fisheries has become increasingly important as discards include not only undesirable species but undersized marketable species. In order to manage fisheries effectively, it is vital that discarding practices of commercial fisheries, particularly of marketable species, are quantified. Stock assessments can then be adjusted accordingly.

Actual measurement of fish discards is difficult and can only be accurately done by having an on-board observer. The expense of this means that only a small proportion of the total fishing trips can be observed. To improve the precision of the estimate of fish discards, the statistical problem may be stated as how to make optimal use of information on a variable x, which is cost effective to measure accurately, in order to estimate the associated variable of interest y, which is difficult or expensive to measure. In this case x may be a measure of vessel characteristics correlated with variable y which is the discard rate or quantity discarded.

2 Method

The relative performance of equal probability estimators (simple random, ratio and regression) and the probability proportional to x (ppx) estimator were evaluated. Data that was observed from Northern Ireland fishing vessels on 35 trips, 21 pelagic and 14 twin-rig trawls during the period April 1997 to August 1998, were investigated to compare the accuracy of the estimates. Additional analysis of information from England, collected from June 1997 to August 1998, and Spain, collected from January to December

1997, enabled validation of the results. Under certain criteria each estimator will prove optimal which will be discussed in relation to the results.

3 Results

For Northern Ireland vessels, there were significant correlations between vessel characteristics and numbers of fish discarded or retained per hour for twin-rig vessels with the regression estimator proving optimal. Substantial improvements, as much as 40%, when comparing the regression with the simple random estimator occurred with a small percentage difference, approximately -5%, between the estimator and the true mean.

There were no significant correlations between vessel characteristics and numbers of fish discarded or retained per hour for Northern Ireland pelagic vessels or any of the English or Spanish vessels. Hence, the simple random estimator proved optimal.

4 Discussion and Further Work

Ppx, regression and ratio estimators use information gathered on x to improve the precision of y, the variable of interest that is to be estimated. For ppx, regression or ratio estimators to be an improvement over the simple random estimate there needs to be a significant correlation between x and y.

Discarding for Northern Ireland could be estimated by a combination of estimators - simple random for pelagic vessels and regression for twin-rig vessels. Due to the differing discarding patterns of the various gear types separate estimators for each gear type within each country could be calculated.

The next collection of discards data is due to commence in July 1999. The additional information will enable a further check of these results. Following on from the results of the last discards project it will be possible to investigate how vessels will be selected for this future project and how to handle the issue of non-response.

Bayesian Prediction with latent Growth Models for On-Line Monitoring in Cheese Manufacture

Rakhi Baj[1] and Bart Mertens[1]

[1] Department of Statistics, Trinity College, Dublin 2, Ireland.

Abstract

There is considerable interest in the Irish dairy industry in the automation of cheese manufacture. This requires the development of prediction methods which can determine the optimum time to cut curd following the addition of rennet to milk. The Department of Agriculture and Food Engineering of University College Dublin has investigated various on-line sensing techniques which can monitor the gelation of the curd. For example, one of the sensors records the change in near infrared reflectance of the curd during coagulation. It is crucial that sensors can only monitor curd formation indirectly.

As in many branches of science, statistics is vital to this problem. Efficient and accurate methods must be developed which can extract and combine information from complex signals recorded during gelation. The objective is the determination of the optimum cutting time. In this project we investigate the specification of latent growth models for curd formation and their application in the modelling of both the observed sensory data and the prediction of the cutting times. It is a particular requirement that the method must be able to account for the considerable seasonal variation in milk composition which is characteristic of Irish milk production. For this reason, we research Bayesian approaches to modelling with latent structure assumptions.

The presentation will introduce the data and problem. Traditional prediction approaches which are currently employed for the analysis of these data are explained and some of the results demonstrated. We then introduce and discuss the Bayesian model and discuss our present results.

Mapping of indoor radon concentration

Yudi Pawitan[1] and Stephen Fennell[2]

[1] Department of Statistics, UCD, Dublin 4, E-mail: yudi@ucd.ie and [2]Radiation Protection Institute of Ireland, Dublin

Abstract

Radon is a naturally occurring radioactive gas that is a known risk factor for lung cancer. It is important to identify areas with high concentration so houses can be built with proper protection to prevent or reduce the indoor radon concentration. ¿From a series of national radon surveys, conducted by the Radiation Protection Institute of Ireland, data were available from 10,106 dwellings in Ireland, located in a total of 758 10km-grid squares. The main objective of the surveys is the estimation of the proportion of houses exceeding a threshold level in each grid square. The standard technique is to analyse the data grid-by-grid separately. The main statistical problem arises from the sparseness of the data in most geographical areas. I will describe a method that uses a parametric model for the distribution within the grid and a nonparametric smoothing of parameter values between neighbouring grids. The two-dimensional smoothing method deals with an irregular object (Irish map), and accounts for varying number of observations (including zero, which means missing grid) and different level of variability between grids.

Keywords: Radon, lung cancer, linear model, smoothing.

1 Introduction

Radon occurs naturally as part of the radioactive decay series of uranium, which is present in rocks and soils. Since radon is an inert gas it moves freely through porous media and rises to the earth surface. In the open air radon is quickly diluted, so it is harmless, but when trapped inside a house the concentration can reach an unhealthy level. The outdoor radon concentration level varies from 4 to 15 becquerels per cubic meter (Bq/m^3). A level higher than 200 Bq/m^3 is considered a health hazard.
The purpose of the national radon surveys conducted by the Radiation Protection Institute of Ireland is to identify areas with high radon concentration. The radon level is usually expressed as the proportion of dwellings

with higher than threshold value of 200 Bq/m^3. An area with a proportion greater than 10% is considered a high radon area. The standard methodology is to estimate the proportion for each 10km-grid separately. The main problem with this method is the sparseness of the data within each grid, or worse: there are grids where there is no data available.
In this talk I will describe the statistical methodology to smooth the radon concentration map in Ireland. The smoothing model (i) takes into account the varying sample sizes and variability of different grids; (ii) allows a naturally irregular shaped map as an input to the model; and (iii) allows a computation of an optimal smoothing parameter.

2 Methods

The country is divided into 10km-grid squares according to the Irish National Grid System, and a random sample of houses are chosen from each grid. Each participating house installed two radon detectors, one for the living area and the other for the bedroom, for a period of one year. The main measurement from each house is a single value, which is average radon concentration over the year and over the two detectors. Data from a total of 10,106 houses are available for analysis; the houses are located in 758 grids, giving an average of 13.3 houses/grid.
Let $y_i(g)$ be the radon concentration of the i'th house in grid g, where $i = 1, \ldots, n_g$, n_g is the number of houses in grid g, the grid index $g = 1, \ldots, G$ and G is the number of grids. We first investigate the distribution of the $y_i(g)$ within a grid. There are many examples where the distribution is log-normal. The log-normal assumption is important to improve the estimation of the threshold probability.
Suppose there is a normalizing transform and let $z_i(g)$ be such transformation of $y_i(g)$, so $z_i(g)$ is normal with mean $\mu(g)$ and variance $\sigma^2(g)$. Let z be a 1-d array of $z_i(g)$'s for all i and g, μ be the array of $\mu(g)$'s such that $E(z) = \mu$. (In this application the length of z is $N \equiv \sum n_g = 10,106$, which the size of the data.) Define a mixed effects model, where conditional on the random effects b we have

$$\mu = \beta_0 + Zb$$

and β_0 is a fixed effect parameter describing the mean level of $z_i(g)$. The factor b is the random field of mean concentration; it is of length G and its elements are denoted by $b(g)$'s. The (j, k) entry of the design matrix Z is one if $k = g$ and zero otherwise, where j is the row associated with a particular $z_i(g)$. The conditional variance of z is $\Sigma = \mathrm{diag}[\sigma^2(g) I_{n_g}]$, where I_{n_g} is $n_g \times n_g$ identity matrix. In effect Σ is diagonal with different values for different grids.

Smoothing is effected by assuming that random effects (or random field) b are positively correlated, with density proportional to

$$\exp[-\frac{1}{2}\lambda \sum_{g_i \sim g_j} \{b(g_i) - b(g_j)\}^2] \tag{1}$$

where $g_i \sim g_j$ indicates grids g_i and g_j are primary neighbours. Under this assumption the log-likelihood of b is of the form

$$-\frac{1}{2}\lambda b' R b,$$

where λR is the inverse covariance matrix implied by (1). Irregular shape in the map is taken into account automatically in the matrix R. The diagonal element $R(g, g)$ is the number of primary neighbours of grid g, and the off-diagonal element $R(g_i, g_j)$ is -1 if $g_i \sim g_j$ and zero otherwise. The parameter λ is a smoothing parameter. Under these assumptions the estimate of b is the minimizer of the penalized $-2\times$log-likelihood function

$$(z - \beta_0 - Zb)'\Sigma^{-1}(z - \beta_0 - Zb) + \lambda b' R b,$$

which is given by

$$\widehat{b} = (Z'\Sigma^{-1}Z + \lambda R)^{-1} Z'\Sigma^{-1}(z - \widehat{\beta}_0), \tag{2}$$

where $\widehat{\beta}_0$ is simply the average of z. To compute $\widehat{b}$ we first need to estimate Σ, which can be done grid-by-grid. A fast solution of (2) is obtained using the Gauss-Seidel method.
An optimal smoothing parameter λ is chosen to minimize an approximately unbiased estimate of $E \parallel \widehat{\mu} - \mu \parallel^2$, where $\widehat{\mu} = \widehat{\beta}_0 + Z\widehat{b}$. Once $\widehat{\mu}$ is found we can immediately compute the threshold probability map, based on the tail of the normal probability.

3 Results

We have found that, after accounting for background concentration, the data are very closely log-normally distributed. This justifies the standard procedure in computing the threshold probability using the normal assumption. Actual procedures have been implemented to compute threshold probabilities based on smoothed mean field.

References

McGarry, et al (1997). *Radon in Dwellings - The National Radon Survey.* Dublin: Radiation Protection Institute of Ireland.

Tree-based Inferences from DNA Sequence Data

David Balding[1]

[1] Department of Applied Statistics, University of Reading, PO Box 240, Reading, RG6 6FN. E-mail: d.j.balding@rdg.ac.uk

Abstract

The analysis of DNA sequence data is going through a period of exciting developments. Genetic samples cannot usually be regarded as random samples: underlying them is a complex pattern of dependencies which are usually best represented by a tree. Traditional methods of inference avoid explicitly modelling the underlying tree, for example by averaging a pairwise statistic over all pairs in the sample. Today, developments in tree-based genetic models and in computational statistics are making feasible a fully likelihood-based approach to inference which explicitly incorporates the dependence structure via a tree. I will review these developments and illustrate them with some inferences about human population histories, mutation processes, and forensic match probabilities.

The Estimation of Numbers of Deaths Attributable To Smoking

Don Barry

[1] University of Limerick Limerick, Ireland E-mail:Don.Barry@ul.ie

Abstract

In a report sponsored by the Imperial Cancer Research Fund (ICRF) and the World Health Organization (WHO), Peto et al (1994) give detailed estimates of smoking - attributed deaths by sex, age, year, and cause for almost 50 developed countries. For instance, of 1529 deaths due to lung cancer in 1990 in Ireland, 1351 are deemed to be attributable to smoking, of 451 deaths due to upper aero-digestive cancer, 265 are attributed to smoking, while of 31,370 deaths due to all causes, 6,417 are attributed to smoking.

The method used to produce such precise estimates of deaths attributable to smoking was introduced in Peto et al (1992). We will describe this method. In order to investigate the appropriateness of this somewhat ad hoc method, we compiled a dataset consisting of the numbers of deaths from each of 12 diseases in the 60-64 age group as well as the size of the population at risk for each of 18 countries and for the year 1990. These data were obtained from the WHO. Across all 18 countries, the total number of deaths among 60-64 year olds from the 12 diseases considered was 151,520 and, of these, 69,557 or 46% are attributed to smoking by the Peto et al methodology.

Suppose that we record the number of deaths from each of D causes in each of C populations. For $i = 1, 2, \ldots, C$, let N_I denote the size of the ith population and for $k = 1, 2, \ldots, D$, let X_{ik} denote the number of deaths due to cause k in population I. Assume that $\{X_{ik} : 1 \leq i \leq C, 1 \leq k \leq D\}$ are independent Poisson random variables with X_{ik} having mean $N_i m_{ik}$. We introduce a G-Category Poisson Model which assumes that

$$m_{ik} = \alpha_I \Sigma_j^G =_{1Pij\lambda jk} .$$

This model has $(C-1)+C(G-1)+DG = G(C+D)-1$ parameters and so only values of G less than $\frac{CD+1}{C+D}$ may be fitted. For $C = 18, D = 12$ this means that the largest value is $G = 7$. We will describe the results of fitting

this model for a variety of values of G. The model with $G = 2$ may be fitted using the Peto et al methodology and we compare those estimates with the estimates obtained by maximum likelihood. The conclusion will be that the Peto et al estimates are very different from the maximum likelihood estimates. We will also conclude that all values of $G \leq 7$ produce models displaying significant lack of fit and report on which countries and diseases contribute most to the lack of fit.

References

Peto R, Lopez AD, Boreham J, Thun M and Heath C, Jr. (1992). Mortality from Smoking in Developed Countries 1950-2000. Indirect Estimates from National Vital Statistics. *Lancet.* No 339, 1268-1278.

Peto R, Lopez AD, Boreham J, Thun M and Heath C, Jr (1994). Mortality from Smoking in Developed Countries 1950-2000: Indirect Estimates from National Vital Statistics. *Oxford University Press, Oxford.*

Numerical State Space Modelling of Binary Data.

John Belcher[1]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: j.belcher@keele.ac.uk

Abstract

Diary data is kept of menstrual bleeding cycles for a group of patients at a local institution. We investigate whether menstrual synchrony occurs in these women, due to the prolonged time they spend together. Menses are recorded as binary data for each women and we use conventional methods, such as Fourier analysis and Spectral Anova techniques (Stoffer, 1991), to investigate menstrual synchrony.

One approach is to consider methods for detecting a common competent of variation in the time series data for the different patients. We call this a *common signal*. Consider $q = 1, \ldots, N$ independent replications of a time series $X_q(t)$ observed at times $t = 0, 1, \ldots, T$. These have a common signal if they are made up of two parts, one, the signal is the same for all the series, the other is independent and different for each series. For a deterministic signal Brillinger [1980] gave the following model specification.

$$X_q(t) = \mu_q + S(t) + \epsilon_q(t)$$

where

(i) μ_q are constants

(ii) S(t) is constant across the patients at any given time t and has power spectrum $F_{ss}(\lambda)$

(iii) $\epsilon_q, q = 1, \ldots, N$ are independent realisations of a stationary time series with mean 0, covariance function $C_{\epsilon\epsilon}(u)$ and power spectrum $f_{\epsilon\epsilon}(\lambda)$.

This approach facilitates the development of a simple test of the null hypothesis of no menstrual synchrony as a function of simple summary statistics calculated from the data.

As an alternative to using these methods we describe how Numerical State Space techniques using Bayes Theorem and time series models can be used

to investigate this phenomenon. This modelling approach will have broader objectives: to develop data generation mechanisms which can successfully mimic (and explain) the observed data, to contain specific parameters which can describe the irregularities of the data, and to have the capacity to be developed to incorporate mechanisms by which the response of one patient is affected by the response of the others.

In a time series context such models will typically incorporate latent variables which underlie the observed data. Such variables are known as state variables, and state transition equations could be developed to describe their evolution through time. The observations could be used to estimate these states as an essential part of the modelling. This framework is much deeper than the simple models which, for example, model the intervals between periods, or length of menses, as random variables. We have attempted to model the data by a simple state variable which may be described as an internal physiological clock.

References

Brillinger, D.R. (1980). Analysis of Variance under Time Series Models. *In Krishnaiah P.R (ed) Handbook of Statistics* Vol. 1, 237-278.

Stoffer, D.S. (1991). Walsh-Fourier Analysis and its Statistical Applications. *Journal of the American Statistical Association.* 86, 461-485.

Fitting Multivariate Survival Distributions To Accident Data For Bus Drivers

Gerry Keane[1] and Don Barry[2]

[1] C.I.E, Heuston Station, Dublin 8. E-mail: gerry.keane@cie.ie and [2]Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland. E-mail: Don.Barry@ul.ie

Abstract

During the period 1992 to 1997, 6,908 accidents were recorded against 2,576 bus drivers in Dublin Bus. The time between accidents for each bus driver is of particular interest. Given the cut-off dates for observation and the date of joining and leaving of bus drivers the data are heavily censored.

A general approach, using copulas and fixed marginals, for constructing bivariate lifetime distributions is put forward. This approach incorporates the traditional frailty model for assessing accident proneness. Using a fully parametric specification of the bivariate distribution, a number of these models are fitted to (t_1, t_2), the time to the first and second accident of each driver respectively. Maximum likelihood techniques are used.

The robustness of the parametric approach is tested by comparison with a non parametric analysis. Because of the nature of the data there is dependent censoring between the t_1 and t_2 data.

Straight forward extensions to the multivariate case are also presented.

Practical Methods for Analysing Dependent Survival Data

John Newell[1]

[1] Department of Statistics, Glasgow University, G12 8QQ, UK. E-mail:johnnew@stats.gla.ac.uk

Abstract

Survival data arises when there is interest in the length of time until a particular event occurs e.g. death due to cancer. Typically observations are assumed to be statistically independent of each other. This assumption however is violated in many situations which are not as uncommon as one might think.

Survival studies involving dependent data fall naturally into two categories namely:

Cluster Studies (where a failure process acts concurrently on individuals in a cluster e.g. time to disease onset in matched studies involving humans, litters etc.) and Multiple Event Studies (where episodes of the same failure process act serially on the same individual e.g. time to exhaustion in repeated exercise testing or time to successive asthma attacks in the same individual).

The aim of this presentation is to illustrate practical methods for analysing dependent survival data primarily from cluster studies.

Keywords: Dependent Survival Data, Cluster Studies .

1 Introduction

Cluster studies are of two types, namely paired studies (e.g. time to cataract in left/right eye) and matched studies where the individuals are matched by design (e.g. comparing time to disease onset for two groups matched on several characteristics). In order to distinguish between these study designs the terms paired survival studies and matched survival studies are used.

Both matched and paired survival studies will have a pair of observation times recorded which represent the two arms of the primary variable of interest. In addition to these, additional information may be recorded also in the form of covariates, or prognostic indicators. Matched survival studies will, by definition, have the variables used for the matching present and some additional unmatched covariates, or prognostic indicators may also be recorded for each individual. Paired studies on the other hand, by definition will not have any matching variables present but may have covariate information recorded for each individual e.g. sex or age.

2 Methods

Graphical techniques for displaying dependent survival data, including bivariate survival scatterplots and survival ratio plots, have been developed. A review of several nonparametric estimators of the bivariate survival function is given with methods for generating reference ranges for such three-dimensional plots. In addition, two methods to graphically assess the independent effect on survival of any continuous covariates are discussed. The first uses a form of kernel estimation to construct an estimator of a percentile of the survivor function as a function of the covariate while the second uses a tree based approach.

In order to make a simple comparison of the survival distributions of the two arms of the primary variable (i.e. ignoring all covariates) a review of several nonparametric paired log-rank tests is given. Two new approaches for comparing survival in paired/matched survival studies are described and illustrated. The first is a simple test of symmetry based on pair performance. The second is based on estimating the distribution of the (pairwise) difference in survival, initially using a parametric approach (for providing an interval estimates of the mean difference in survival time) and finally using a nonparametric approach (where inference can be centered on the appropriate quantile e.g. the median difference).

Methods for incorporating covariates into the analysis, while at the same time taking the dependency structure of the data into account, are presented. A comparison of the two arms of the primary variable should be less biased and more precise than a simple comparison. In matched survival studies the matching covariates are available for inclusion in the analysis while in paired studies the covariates representing the degree of similarity for the pair are often unobservable, that is, hidden from the analysis.

A new approach for modelling pair performance which allows for covariates is presented. Regression models for the hazard rate are discussed. Several extensions of the proportional hazards (PH) model to clustered studies

are proposed. The conditional PH model ignores the matched structure of the data, however it uses information on the matching to correct inference made on the primary variable. The justification is that the model assumes conditional independence by forcing in the matching covariates in the final model. A different approach is to model the data, initially ignoring dependency, producing the standard estimates of the regression parameters. The regression coefficients are estimated assuming independence while the estimated covariance matrix is then corrected post fit using a paired-jacknife estimate of the variance.

A further refinement to the PH model for paired/matched survival data is to allow each pair to define a separate stratum. The association within each pair is then considered a fixed effect. An alternative more elaborate procedure introduces a random term for each pair that represents the within-pair association. In a final extension to the PH model a random term corresponding to each pair is introduced into the model. This random pair effect, often termed a frailty, generates dependency between the survival times of the individuals in a pair. The random effects represent unobserved covariates. Random effects are considered to act multiplicatively on the individuals hazard rate. Survival times of all individuals are then assumed to be independent given random effects (and any observed covariates).

3 Results

The results of a large simulation study which compare the different methods proposed for analysing dependent survival data are given. A range of different degrees of censoring, sample size and effect size combinations are covered. Illustrations of these methods are presented using paired survival data from an Orthodontic study and matched survival data from a Melanoma study.

Predicting The Survival Of Stroke Patients In Hospital Using Bayesian Belief Networks And Phase-Type Distributions

Adele Marshall[1], Sally McClean[1], Mary Shapcott[1] and Peter Millard[2]

[1] School of Information and Software Engineering, University of Ulster, Jordanstown Northern Ireland, BT37 0QB and [2] Department of Geriatric Medicine, St George's Hospital, London, SW17 0RE, UK.

Emails: AH.Marshall@ulst.ac.uk, SI.McClean@ulst.ac.uk, CM.Shapcott@ulst.ac.uk and P.Millard@sghms.ac.uk

Abstract

In this paper, we introduce Dynamic Bayesian Belief networks which generalise the concept of Bayesian Belief Models to include a duration of time. We may thus represent a stochastic process along with causal information and an outcome. Structured Phase-type (Ph) distributions which characterise a type of Markov model are here used to provide an intuitive and robust way of describing such probabilistic processes. Such models describe duration until an event occurs in terms of a process consisting of a sequence of phases - the states of a Markov model. The outcome is then one of a number of possible ways in which the process may terminate. This is similar to survival analysis which models duration until a particular event occurs. One feature considered by survival analysis is the inclusion of missing or censored data. In our case, we overcome this problem of censored data by estimating survival using the EM algorithm. Our approach is illustrated using data on hospital spells (the process) for a number of geriatric patients along with information on whether or not the patients suffered a stroke immediately prior to admission to hospital (the cause) and whether they were alive or dead on discharge from hospital (the outcomes).

Keywords: Bayesian belief networks, Phase-type distributions, Dynamic Bayesian belief networks.

1 Introduction

A Bayesian belief network (BBN) is a graphical representation of uncertain knowledge that most people find easy to construct and interpret. In addition the representation is probabilistic (Heckerman, 1996, Cox and Wermuth, 1996). One of the key features of BBNs is the ability to represent causality.

Phase-type distributions are a special type of Markov model (Neuts, 1981). They describe the time to absorption of a finite Markov chain in continuous time, where there is a single absorbing state and the stochastic process starts in a transient state (Faddy and McClean, 1999). Phase-type distributions can be generalised to include almost continuous distributions (Faddy, 1994) therefore making them appealing to use.

In this paper, we introduce Dynamic Bayesian belief networks (DBBNs) which generalise the concept of Bayesian belief models to include a duration of time. We may thus represent a stochastic process using phase-type distributions along with causal information, represented by a BBN, and an outcome. For example, we may have a BBN with causal nodes: 'source of referral', 'diagnosis' prior to the event 'admission to hospital' which initiates the process 'stay in hospital'. This DBBN may then have the associated outcome node 'discharged dead/alive'.

The application considered, is the planning and management of a particular group of geriatric stroke patients in hospital. As the proportion of the elderly in the population continues to rise, geriatric departments are faced with difficult decision on how to effectively allocate resources. The knowledge of the length of stay in a patient in hospital would be invaluable to the management of the geriatric department.

2 Model

We define the DBBNs as comprising causal nodes $\mathbf{C} = \{C_1, \ldots C_m\}$, process nodes $\mathbf{Ph}\{Ph_1, \ldots Ph_n\}$ and an outcome node $\mathbf{O}$ (Figure 1). The parameters of the process (the phase-type distribution) are conditional on the outcome. For our example, we would thus fit two different phase type distributions: one for patients who died in hospital and one for patients who were discharged alive. We may therefore represent the joint distribution of $\mathbf{C}$ by:

$$P(\mathbf{C}) = \prod_i P(C_i \mid pa(C_i))$$

where pa is the parent set of C_i.

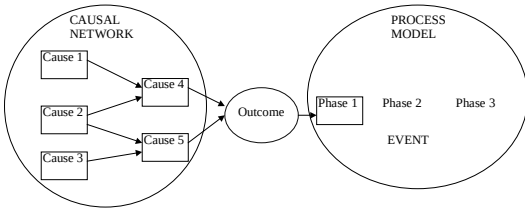

FIGURE 1. The Dynamic Bayesian Belief Network with an outcome node

3 Application to Geriatric Medicine

The data has been taken from 'CLINICS', a clinical computer system in use in elderly care at St George's Hospital, London. The admission details, length of stay in hospital and the outcome of each patient in the geriatric department were recorded. This analysis is concerned with predicting the length of stay for stroke patients. The network presented in Figure 2 describes the length of 'stay in hospital' for stroke and non-stroke patients by two different two term mixed exponential distributions. We have one causal node (stroke?), one outcome node (discharged, alive or dead) and one phase for the 'stay in hospital' process.

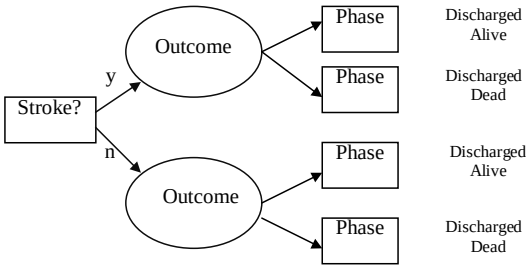

FIGURE 2. The Geriatric DBBN

We found that the length of stay in hospital for stroke patients can be described by the following two term mixed exponential distribution:-

$$f(t) = 0.6797e^{-0.0196t} + 0.3203e^{-0.0263t} \quad (1)$$

This was used to predict the (expected length of stay which was compared with the actual observed data by performing a chi-square (χ^2) test. The results demonstrate that the two term mixed exponential taken from the network gives a good (99.9%) representation of the length of stay of the stroke patients with outcome 'discharged alive'. The chi-square value for

stroke patients 'discharged dead' shows that the model is also a good representation for those patients who die in hospital.

4 Summary

We have introduced a particular type of latent Markov model - the conditional Phase-type distribution - to describe local dependencies with respect to representation of a stochastic process. Our resulting Dynamic Bayesian Belief Network is hybrid, in that we use discrete variables for the causal model and a continuous variable for the stochastic process. We have also addressed the issue of censored data by using the EM algorithm to estimate the survival distribution. The application we have considered has proven to be a good representation of the original survival data.

References

Cox D.R and Wermuth N (1996). Multivariate dependencies. *Chapman and Hall.*

Faddy M (1994). Examples of fitting structured phase-type distributions. *Applied Stochastic Models and Data Analysis.* 10, 247-255.

Faddy M and McClean S.I (1999). Analysing Data on lengths of Stay of Hospital Patients Using Phase-Type Distributions. *Applied Stochastic Models and Data Analysis.* To appear.

Heckerman D (1996). Bayesian networks for Knowledge Discovery. In Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R (eds) (1996). *Advances in Knowledge Discovery and Data Mining AAAI Press/The MIT Press* 273-305

Neuts M (1981). Matrix Geometric Solutions in Stochastic Models. *John Hopkins University Press, Baltimore, Maryland.*

Survival Analysis for Longitudinal Data?

Gilbert MacKenzie[1]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: g.mackenzie@keele.ac.uk

Abstract

In longitudinal studies with a set of continuous or ordinal repeated response variables it may be convenient to summarise the outcome as a threshold event. Then, the time to this event becomes of interest. This is particularly true of recent Ophthalmological trials evaluating the effect of treatment on the loss of visual acuity over time. However, the practice of employing conventional survival analysis methods for testing the null hypothesis of no treatment effect in these types of studies is intrinsically flawed as the exact time to the threshold event is not measured. In this paper we obtain a general Likelihood for the unknown parameters when the underlying survival model is parametric. We also recover the actual information available in repeated measures data for a variety of models and compare the results with those obtained using a mis-specified model, which assumes the time to the event is one of the possibly irregularly spaced inspection times.

Keywords: Longitudinal Data, Survival Analysis, Model Mis-specification, Grouped Likelihood.

1 Introduction

In longitudinal studies in which the response is continuous or ordinal, clinicians often find it convenient to categorise the outcome. If the response is change from baseline it may be natural to think in term of a threshold effect. Opthalmological studies usually investigate visual loss over time in terms of distance visual acuity (DVA) measured on an ordinal scale due to Bailey-Lovie (B-L) and analyses are frequently performed in terms of numbers of lines of visual acuity lost (MPSG, 1994). For example, Bergink et al (1998) in their RCT of teletherapy in ARMD analysed the outcome <3 or $3+$ lines of visual acuity lost. In these studies recruitment is staggered over time and increasingly survival-type methods, such as Kaplan Meier curves, the Log-Rank test, and the Proportional Hazards model of Cox (1972) are being pressed into service.

These methods are appropriate for right censored 'time to event data' when the exact time of occurrence is known, but strictly inappropriate when the 'time to event' is known only to lie in an interval. Thus, for example, the loss of 3+ BL lines observed at a scheduled inspection time could have occurred at any time since the previous examination. Thus, the use of inspection times to rank the data is merely approximate. It is obvious that this procedure must lead to many ties in the so-called times to the events. Such ties are often ignored or inappropriately broken by the variation in examination times, rather than by the actual times when the threshold was crossed.

In this paper we (a) identify the correct Likelihood for pseudo 'time to event data' arising in repeated measures studies (b) obtain the maximum Likelihood estimating equations for some parametric models and (c) compare inference under the correct model with the mis-specified model, which utilises the inspection times as if they were exact.

References

Bergink, et al (1998). A Randomised Controlled Clinical Trial on the efficacy of radiation therapy in the control of subfoveal choroidal neovascularisation in age-related macular degeneration: radiation versus observation. *Graefe's Arch. Clin. Exp.Ophthalmology.* 236, No 752, 1-5.

Macular Photocoagulation Study Group. (1994). Visual outcome after laser photocoagulation for sub-foveal choroidal neovascular secondary to age-related macular degeneration. *Arch. Ophthalmol.* 112, 480-488.

Cox DR (1972). Regression models and life tables (with Discussion) *JRSS B.* 34, 187-220.

MacKenzie G (1996). Regression models for survival data. *JRSS D.* 45, 1, 21-34.

MacKenzie G (1997). On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine.* 16, 1831-1843, 21-34.

Reeves J and MacKenzie G (1998). A bivariate regression model with serial correlation. *JRSS D.* 47, 4, 607-615.

Joined-up Government and Joined-up Statistics

John Fox[1]

[1] Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ

Abstract

There is increased rhetoric about joined up government. What will this mean in practice and what are the implications for social statistics? The "Our Healthier Nation" strategy will be used to explore the main implications.

The Government Statistical Service aims "to inform debate decision-making and research both within government and by the wider community". ONS has set the following success criteria: to improve quality and relevance of the service to customers; to minimise the burden on those supplying the information; to improve public confidence in the integrity and validity of our outputs; to improve value for money; and to maintain a well motivated workforce.

Recent developments in national statistics show what progress is being made to achieve the long term vision and what remains to be done. Clear progress has been made in terms of presentation, communication with stakeholders, and in the introduction of improved data but we are expected to do better particularly in terms of further efficiency gains to free up additional funds for new investment and in terms of the way we quality assure, research and introduce new methods.

The 2001 Census will introduce a number of major innovations and will provide the basis for our building a more comprehensive picture of social change for 2001 and beyond. There will be other opportunities in the first decade and these will enable us to respond more effectively to the challenges of joined up government. Some of these will be discussed to stimulate thoughts on what early thinking we need now in order to take full advantage of them."

Statistics and Public Questions

Stephen Haslett[1]

[1] Statistics Research and Consultancy Centre, Massey University, PO box 11222, Palmerston North, New Zealand

Abstract

In a number of countries there has been a longstanding interplay between Statistics and questions of public interest, where the consequent debate involves (or ought to involve) the veracity, relevance, completeness, or interpretation of collected data.

The formation and existence of a national Statistical Association provides a possible vehicle for formalising this process by providing a neutral body, and thus, via a constituted committee, a mechanism for placing a clear appraisal of the statistical aspects of the debate into the public arena.

The decision by a professional body of statisticians to form such a committee, and the protocol by which decisions are made on how the committee should be involved in individual issues, are important matters. At a finer level, so are the process by which necessary formal structures to undertake an appraisal instituted, appraisals are undertaken, and decisions on the extent of that involvement made. While many of these matters are in principle statistical, care in practice is often required to avoid what may later be interpreted by interested parties as political or commercial bias.

This paper discusses both these general issues and the case of the New Zealand Statistical Association, founded in 1948, and its Survey Appraisals and Public Questions Committee (SAPQC).

The SAPQC was set up during the 1970's with the following objectives:

"To raise the standard of practice and the level of public understanding of statistics in New Zealand by:

(a) conducting independent appraisals of sample surveys, opinion polls and other statistical statements in relation to their statistical validity, and to the needs of the users of the survey results;

(b) conducting examinations of statements made in the public domain and of significant public interest, that have statistical content, or whose validity depends on statistical considerations."

This paper provides some history of the SAPQC, which the presenter has convened for the last ten years. The paper discusses the brief, protocols, and raison d'etre of this committee, as well as the benefits and disadvantages of such a committee. The general issues and principles are outlined by reference to some of the major appraisals undertaken since 1992.

References

(May 1992). Comments on "A New Zealand Poverty Measure". *Report prepared by BERL, New Zealand Statistics, and the Institute of Policy Studies.*

(1992). Appraisal of the Commerce Commission's Telecommunications Industry Inquiry Report.

(July 1992). Comments on New Zealand Valuation Procedures.

(July 1992). Submission to the Review of Fisheries Research.

(December 1992). Appraisal Energy Direct's Survey to Establish Preferred ownership.

(February 1993). Appraisal of Waitemata Electricity's Surveys to Establish Preferred Ownership.

(May 1993). Appraisal of MRL Research Group's Surveys to Establish Preferred Ownerhsip of Central Power.

(September 1994). Report on the Statistical Adequacy of Current Monitoring of Social Welfare Benefit Levels.

(September 1994). Submission on Ministry of Health Initiative "New Avenues for Crown Funded Social Science Research" especially Proposal 4: 'Establishment of a Social Science Research Clearing House'.

(June 1995) Appraisal of Two BERL Reports on Cost Benefits of Section 21 of the Marine Transport Act 1994.

(December 1995) Appraisal of the Second "Evening Post" Survey on Voter Preference for Candidates in the Wellington Mayoral Election 1995

(May 1998) Appraisal of "Towards a Code of Social and Family Responsibility: Public Discussion Document, February 1998".

On Measuring Industry Compliance with Environmental Laws

Richard Bolstein[1]

[1] Department of Applied and Engineering Statistics, George Mason University, Fairfax, VA 22030, USA.

Abstract

The Government Performance and Results Act (GPRA) passed by the United States in 1993 requires federal agencies to periodically measure and report on performance. Up to this time, agencies measured performance by outputs: activities they controlled such as the number of inspections, number of violations found, and the number of enforcement actions taken. GPRA now requires agencies to focus on outcomes: activities they influence. A key outcome measure is the rate of compliance or noncompliance with a government regulation. It is presumed that agency efforts should lead to an improvement in the compliance rate over time. The problem is how to define meaningful rates that are computationally feasible and for which statistically valid estimates are possible.

This paper focuses on work in progress at the Environmental Protection Agency (EPA) of the U.S to measure industrial compliance with statutes of the Clean Water Act, Clean Air Act, and Toxic Waste Disposal Act. Depending on the industry and relevant statute, some facilities are required to submit periodic self-reports on the level of release of toxic substances into the environment, while others are periodically inspected either by state or federal agents. The problems in defining and computing statistically valid compliance rates include:

1. Self-reports have potential for bias.
2. There is a high non-response rate in self-reporting populations.
3. Inspections are not conducted at random.
4. Inspections usually occur on a single day and may not detect violations that occur at other times during the period (month/quarter/year) for which the compliance rate is desired.

These problems cannot be fixed. For example, inspections are usually targeted to facilities suspected of non-compliance, and are inspected at expedient, rather than random times. Since there is still pressure to catch violators, there is natural resistance to random sampling. Consequently, methods must be found to make use of exisiting data and supplement it with random sampling in order to achieve statistical validity. This paper presents our initial ideas on how to define and compute statistically valid compliance rates for self-reporting and inspected populations.

How Do Recruitment and Admission Affect Official Estimates of the Length of 'Time-to-Admission'?

Paul W Armstrong[1]

[1] Department of Health Sciences, University of East London, Romford Road, London, E15 4LZ. Email:P.W.Armstrong@uel.ac.uk

Abstract

In the UK, the Government Statistical Service reports the percentage of elective admissions that took place within three months of a patient being added to the waiting list. This percentage is calculated from cross-sectional data using the total number of elective episodes within a specified calendar period as denominator and the number of these enrolled on the waiting list less than three months previously as numerator. The Government Statistical Service publishes this statistic as a measure of the likelihood of elective admission within three months of recruitment.

Now the number of elective admissions within 0-3 months reflects the likelihood of admission and the numbers 'at-risk' of admission within the waiting time category and calendar period of interest. In other words, the number of elective admissions within the 0-3 month waiting time category will increase if there is any increase in the likelihood of admission or in the size of the population exposed to that likelihood. So the numerator used by the Government Statistical Service reflects conditions within the waiting time category throughout the period of interest.

The admissions observed in each waiting time category are added together to give an indication of the overall size of the population eligible for elective admission ie, the data is handled as though it belonged to a cohort followed to extinction rather than a cross-sectional snap-shot. This total assumes that the number of patients eligible for elective admission 3-6 months after enrolment is identical to the number surviving admission from the 0-3 month category although the two groups belong to cohorts of patients which were recruited quite independently. In other words, the existing approach views the waiting list as a closed and stationary population and only provides an unbiased estimate under these conditions.

If waiting lists are not stationary, we should expect the Government Statistical Service method to produce biased estimates of the likelihood of admission. This paper explores the effect of relaxing stationary population assumptions to the extent apparent in Department of Health data.

Acknowledgements

This work was conducted while the author was registered as a PhD student at the London School of Hygiene and Tropical Medicine, sponsored by South-Thames R&D Directorate and based at Kingston & Richmond Health Authority.

Social Deprivation Indicators and the Geographical Distribution of Opiate Use in Young Males in Dublin

C.M. Comiskey[1]

[1] Mathematics Department, National University of Ireland, Maynooth, Co. Kildare, Ireland

Abstract

Comiskey (1998) provided the first estimates of known and hidden opiate use in Dublin and in so doing identified areas of Dublin with high prevalence. Following this analysis, a regression and correlation analysis of spatial distribution of known drug use amongst young males and material deprivation was performed. The SAHRU (1997) national material deprivation index was used. In order to identify which factors contributed most to prevalence further detailed census data on social indicators including level of employment, social class and level of education was obtained and correlated with prevalence.

Known prevalence for each postal district in Dublin was plotted against mean deprivation index for that district. Several regression models were fitted, these included linear, quadratic, exponential and quadratic exponential with the quadratic exponential providing the best fit. In each case a strong correlation was found between known prevalence per 1000 of males aged between 15 and 24 years in each postal district and material deprivation ($r = 0.62$ for the quadratic exponential model).

In order to assess which indicators contributed most to the prevalence of opiate use additional detailed census data was obtained on the indicators and the key social factors contributing to prevalence were identified. These results will enable policy planners to allocate resources and target those factors which are contributing most to the use of opiates among young males.

Aggregating Uncertain and Imprecise Information for Data Mining

Sally McClean[1], Bryan Scotney[1], and Mary Shapcott[1]

[1] School of Information and Software Engineering, Faculty of Informatics, University of Ulster, Cromore Road, Coleraine, BT52 1SA, Northern Ireland

Abstract

We do not know the exact values of uncertain data, but rather are provided with a probability distribution; this may be because of the nature of data collection e.g. from distributed databases. Imprecise data, on the other hand, occurs because we are not certain about the specific value of an attribute (variable) but only that it takes one of a group of possible values. We here consider the problem of aggregating data which is subject to both uncertainty and imprecision. The method of aggregation uses the Kullback-Leibler information divergence between the aggregated probability distribution and the individual data values; this is here equivalent to maximising the likelihood. We are thus able to use such data to provide a probability distribution for the values of a variable or group of variables.

1 Background

Frequently, real life data are uncertain, i.e. we are not certain about the truth of an attribute value, or imprecise, i.e. we are not certain about the specific value of an attribute but only that it takes one of a group of possible values. Such imprecision might occur naturally as a result of data being provided at different classification levels. A recent survey of various approaches to handling such information in Data and Knowledge Bases has been provided by Parsons (1996).

In this paper we consider the problem of aggregation for such a data model. Whilst traditional query processing is tuple(case)-specific, where we need to extract individual tuples of interest, processing of uncertain data is often attribute(variable)-driven where we need to use aggregation operators to discover properties of attributes of interest. Thus we might want to aggregate over individual tuples to provide summaries which describe relationships between attributes. Such a facility is a central requirement in

providing a database with the capability to perform the operations necessary for Data Mining, where we are frequently concerned with identifying interesting attributes or interesting relationships between attributes.

2 The Approach

A suitable data model has previously been proposed by Barbar et al. (1992) who introduced the term probability data model (PDM). We define a general aggregation operator which minimises the Kullback-Leibler information divergence between the aggregated probability distribution and the data. This is equivalent to maximising the likelihood of the model given the data. It generalises similar operators for crisp data in a manner which apportions uncertain belief in an intuitive manner. Vardi and Lee (1993) have shown that this problem belongs to a general class which they term Linear Inverse Problems (LININPOS). For such problems they develop an iterative scheme which is then shown to converge monotonically to the solution of the minimum information divergence equation. This iterative scheme is in fact an example of the EM (expectation-maximisation) algorithm (Dempster et al, 1977), which is widely used for the solution of incomplete data problems. For example, Maximum Likelihood Estimation has been used for the integration of distributed data over partitions of an attribute domain (Scotney and McClean, 1999).

Whilst it is useful to compute such aggregates, it is often of more interest to derive the joint probabilities of several attributes' values. This is particularly the case for knowledge discovery where, for example, we might want to investigate whether the probability of heart disease is higher for smokers than for non-smokers. Our methodology may be extended to cover the situation in which we wish to compute aggregates of several variables.

3 Knowledge Discovery

In order to discover knowledge from databases we are often concerned with inducing beliefs or rules from database tuples (Anand et al., 1997). Rules tend to be based on sets of attribute values partitioned into an antecedent and a consequent. A typical "if then" rule of the form "if antecedent = true, then consequent = true" is given by "if an individual smokes and is hypertensive, then he or she is very likely to suffer from heart disease". Support for such a rule is based on the proportion of tuples in the database that have the specified attribute values in both the antecedent and the consequent. Using our approach we may use the aggregation operator to derive joint probabilities and then compute conditional probabilities to assess the potential for new knowledge. By using such an approach we may therefore determine the strength and support for various rules under consideration.

A probabilistic approach, such as we have described, may then be used to represent rules as linguistic summaries (McClean and Scotney, 1997) in order to provide a user friendly interface to the Data Mining process.

4 Acknowledgements

This work was partially funded by IDARESA (ESPRIT project no. 20478) and partially funded by ADDSIA (ESPRIT project no. 22950) which are both part of EUROSTAT's DOSIS (Development of Statistical Information Systems) initiative.

References

Anand S.S, Scotney B.W, Tan M.G, McClean S.I.*et al* (1997). Designing a Kernel for Data Mining. *IEEE Expert.* 12, (2) 65-74.

Barbara D, Garcia-Molina H and D Porter (1992). The Management of Probabilistic Data. *IEEE Transactions on Knowledge and Data Engineering.* 4, 487-501.

Dempster A.P, Laired N.M and D.B Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorith (with discussion). *J.R. Statist.Soc.B.* 39, 1-38.

McClean S.I, and B.W Scotney (1997). Using Evidence Theory for the Integration of Distributed Databases. *International Journal of Intelligent Systems.* 12, (10) 763-776.

Parsons S (1996) Current Approaches to Handling Impferfect Information in Data and Knowledge Bases. *IEEE Transactions on Knowledge and Data Engineering.* 8, 353-372.

Scotney B.W, McClean S.I and M.C Rodgers (1999). Optimal and Efficient Integration of Heterogenous Summary Tables in a Distributed Database. *Data and Knowledge Engineering Journal* Forthcoming.

Vardi Y and D Lee (1993) From Image Deblurring to Optimal Investments: Maximum Likelihood Solutions for Positive Linear Inverse Problems (with discussion) *J.R Statist. Soc. B.* 569-612.

Towards Metadata-Guided Distributed Statistical Data Processing

Ronan Pairceir[1], S.I.McClean[1], W. Grossmann[2] and K.A Froeschl[2]

[1] School of Information and Software Engineering, University of Ulster,Coleraine, Northern Ireland and [2] Department of Statistics, Operations Research and Computer Methods, University of Vienna, Austria

Emails: rpairceir@causeway.infc.ulst.ac.uk, si.mcclean@ulst.ac.uk, grossmann@smc.univie.ac.at, froeschl@smc.univie.ac.at

Abstract

Statistical databases routinely handle information about aggregated *macro-data* as well as conventional microdata. In addition to storing macrodata, modern statistical databases usually hold a considerable amount of accompanying metadata. In addition the data may be held in a distributed environment. This paper describes the statistical process along with data models and operators which have been developed as part of an Internet-based metadata-guided distributed statistical processing system which emanates from a DOSIS project called i_{daresa} - an Integrated Documentation and Retrieval Environment for Statistical Aggregates. The objective of i_{daresa} is to provide users (statistical data consumers) with online service via the Internet. We discuss the implementation of the distributed statistical application environment where distribution has been carried out via the Internet.

1 Introduction

The evolution of database technology has resulted in the development of efficient tools for manipulating and integrating large amounts of data. Frequently these data are distributed among different computing systems on various sites. Distributed Database Management Systems provide a superstructure which integrates either homogeneous or heterogeneous DBMS (Bell and Grimson, 1992). In recent years, there has been a convergence between Database Technology and Statistics. The major statistical packages

now incorporate some database functionality, while there has been increasing pressure for statistical databases to be developed which include all the advantages of modern Database Technology. In Europe, this development has been particularly encouraged by DOSIS, which is part of the ESPRIT Programme in the EU Framework IV initiative. DOSIS, which is being carried out in association with EUROSTAT - the EU statistical agency, supports European projects concerned with the development of statistical information systems.

In this paper our focus is on the statistical process which we describe in terms of atomic operators which are chained together to provide automated Metadata-Guided Data Processing. We here describe the process in general terms without tying it to specific solutions although some details of our implementation are also provided.

2 Methods

Statistical databases have traditionally handled information about micro (raw) data and/or macro (summary) data. When such statistical databases are distributed among computing facilities at various sites a new dimension is added to the problems of the database designer who is now faced with much more complicated issues of data processing and perhaps transaction management as well as having to devise ways of integrating data which are heterogeneous in various respects. In addition to storing microdata and macrodata, modern statistical databases usually hold a considerable amount of accompanying metadata (Grossmann 1996, Froeschl 1997). In the i_{daresa} DOSIS project the data along with its associated metadata is known as a tandem object. The statistical application environment which is being used to implement the tandem objects is based on the MIMAD (micro/macro) data model, previously developed at the University of Ulster (Sadreddini et al. 1990, 1992a, 1992b, 1992c). However, this approach has needed to be extended to take account of metadata and to interface with various other i_{daresa} data objects.

Having defined the tandem model for macrodata and microdata along with corresponding models for the accompanying metadata, a number of database operations are also required. Using these operations a user can define a final table which is obtained after some processing steps. This final table may be obtained by means of a simple operation or may require several intermediate derived tandems, or views. Such intermediate views may then be integrated and processed in a final step to give the result. In this case, at each intermediate step, the metadata must be transformed as well as the macrodata. This process has been described in Grossmann and Froeschl (1997) as $Metadata - GuidedDataProcessing$.

3 Resulting Statistical Operators

Statistical data may be made available either as microdata, which provides information on individuals or statistical units, and/or macrodata which are statistical summaries. In either case, the statistical user is primarily interested in producing summarised information and so, early in the process, the microdata is converted to a *summarised* macrodata format (the Σ step). Prior to this, however, some data *preparation* may be carried out on the microdata (the P step). A first step in macrodata processing usually involves some form of data *selection* (the S step) where we may select only a subset of possible variables or may define new variables known as **statistical domain concepts**. In a distributed environment the data must then be *harmonised* (the H step) prior to *fusion* (the F step). The final data processing step is data *presentation* (the π step) where the data are converted into a form suitable for presentation to the user. The statistical process is illustrated in Figure 1.

We denote micro-, macro- and metadata by μ, M and m and micro and macro tandems by τ and T respectively where τ is a function of micro- and metadata and T is a function of macro- and metadata. Then:

$$P : \tau_1(\mu_1, m_1) \rightarrow \tau_2(\mu_2, m_2); \qquad \Sigma : \tau_1(\mu_1, m_1) \rightarrow T_2(M_2, m_2);$$
$$S : T_1(M_1, m_1) \rightarrow T_2(M_2, m_2); \qquad H : T_1(M_1, m_1) \rightarrow T_2(M_2, m_2);$$
$$F : \{T_i(M_i, m_i)\} \rightarrow T_2(M_2, m_2); \quad \pi : T_1(M_1, m_1) \rightarrow T_2(M_2, m_2).$$

A statistical query may then be specified in terms of operators and tandems expressed as nested function calls. An example of such a query specification is:

$$\pi(F\{H(s(T_1(\mu_1, m_1)), H(s(T_2(M_2, m2))\})$$

where we select and harmonise the macro tandems T_1 and T_2 respectively, fuse the result and then present it to the user.

4 Implementation

In our prototype the tables comprising the micro-, macro- and metadata, as well as the corresponding result tables after query processing, are stored in a SQL Server. Here the query process is the execution of SQL query statements. Access to remote servers is achieved via the Internet in a Java environment. A well acknowledged three tier architecture has been adopted for the design. The logical structure of the architecture consists of a front-end user (the client), a back-end user (the server), and middle-ware which maintains communication between the client and the server. We have used

a distributed computing middleware capability called remote method invocation (RMI) which is built into Java.

Thus far we have demonstrated basic functionality of the tandem operators along with the associated mechanisms for remote execution of queries. However, optimisation issues have yet to be addressed, as far as the operators are concerned, although in principle the remote access of summary data is intrinsically efficient since we can regard such data as a compressed version of the micro data.

References

Bell D.A and Grimson J.B (1992). Distributed Database Systems. *Addison Wesley.* 410.

Froeschl K.A (1997) Metadata Management in Statistical Information Processing *Springer Wien, New York*

Grossman W and Froeschl, K.A (1997). IDARESA - A Study in Practical SMIS Design. *Submitted to METIs '98.*

Sadreddni M.H, Bell D.A and McClean S. (1990) Architectural Considerations for Providing Statistical Analysis of Distributed Data *Information and Software Technology* 32, 459-469.

Sadreddni M.H, Bell D.A and McClean S.I (1992a). A Framework for Query Optimisation in Distributed Statistical Database. *Information and Software Technology.* Vol 34. No 6.

Sadreddni M.H, Bell D.A and McClean S. (1992b). A Model for Integration of Raw Data and Aggregate Views in Heterogenous Statistical Databases. *Database Technology* Vol 4, part 2, April, pp 115-127.

Sadreddni M.H, Bell D.A and McClean S. (1992c). Providing Statistical Functionality in a Distributed Environment in Westlake, A Barks, R Payne, C Orchard, T. *Survey and Statistical Computing* North-Holland, September, pp 467-476.

Promoting Statistical Discovery and Retrieval Using Statistical Metadata

Y. Bi[1] and F. Murtagh[2]

[1] School of Information and Software Engineering, Faculty of Informatics, University of Ulster, Shore Road, Newtownabbey, Co Antrim, BT37 0QB and [2] School of Computer Science, Queen's University of Belfast, Belfast, BT7 1NN, Northern Ireland

Abstract

Immediate access to statistical information has become more than just a desire. The network information systems supporting such access have steadily improved as the underlying World Wide Web infrastructure has improved. However, in response to this, the rich statistical store of information, which is rapidly growing and is widely distributed across the Internet, presents a range of new problems for effectively and efficiently discovering and accessing statistical resources for the handling and analysis of statistical data. These problems challenge conventional methodologies and technologies in information retrieval, database management, and statistical handling and analysis. The ESPRIT ADDSIA (Access to Distributed Databases for Statistical Information and Analysis) project is seeking effective approaches to these problems. One theme of the project is how to use statistical metadata to organise statistical information resources across the Statistical Offices in Europe, for promoting statistical information search and discovery.

Statistical metadata is a key term in statistical information, which is used to refer to the characterization of statistical data objects for purposes of locating, accessing and evaluating the distributed sets of statistical data objects. In the ADDSIA project, statistical metadata is divided into the two categories of browsing metadata and processing metadata. The former is used for the description and discovery of statistical information. The latter involves the handling and analysis of statistical data incorporating statistical marcodata and microdata. At issue here is essentially the first category, browsing metadata.

The typical approach adopted for statistical metadata involves the selecting of representations of concepts that characterize the structure, context, and content, and links with networked information resources; the extraction of such representations from free-text documents or heterogeneous statistical

information; and the use of these representations for statistical resource discovery and access which includes organizing, indexing, and searching.

The Document Type Definition (DTD) of eXtensible Markup Language (XML) provides a robust facility for the specification of metadata and the description of documents using such metadata. XML focuses on the content description, rather than on the display of the document. The structure of an XML-coded document is formally defined in a set of markup declarations which constitutes a content model - DTD. These declarations describe a set of markup instructions, known as tags, which can be used to identify the start or the end of logically constituted parts in the XML documents, and the links with external information resources. In addition, there is a semantic relationship between the tags and the parts demarcated by the tags.

In this paper, firstly, we explore how to use the statistical metadata to represent the statistical information - which are possibly complex data objects - textual or documentary, numeric flat tables, relational databases, and general document figures. Secondly, we describe the specification of a DTD for statistical data and metadata, and address how to impose a structure of metadata on unstructured statistical data. Thirdly, we give an architecture for the Metadata Module in the ADDSIA project. Finally, we develop a prototype for statistical information resource discovery using statistical metadata, which has the following features:

- Comprehensive support for description of domain content using XML
- Using agreed metadata, it is straightforward to define a markup language
- A first cut at such a language for economic indicators has been made
- Support for many types of data objects–text, numeric, graphical, etc.
- Display is clearly separated from other content-related operations (indexing and search).

References

Ora Lassila (1998). Resource Description Framework (RDF) Model and Syntax. *W3C Working Draft.*

Andrew Layman, Edward Jung, Eve Maler and Henry S Thompson (Jan 1998). XML-Data. *W3C.*

Stuart Weibel and Juha Hakala (February 1998). A Report on the Workshop and Subsequent Developments. *D-Lib Magazine.* ISSN 1082-9873.

Moen and McClure (1997) An Evaluation of U.S *GILS Implementation*

Stuart Weibel and Renato Iannella (March 1997) The 4th Dublin Core Metadata Workshop Report. *DC-4. National Library of Australia, Canberra.*

Martin Bryan (1997) SGML and HTML Explained. *Addison Wesley Lonhman.*

Lincoln D Stein 1997 How to Set Up and Maintain Web Site. *Addison-Wesley*

Len Seligman and Arnon Rosenthal A Metadata Resource to Promote Data Integration. *IEEE Metadata Conference.*

Recent Developments in the Preparation for the 2001 Census in Northern Ireland

Robert Beatty[1] and Maire Rodgers[1]

[1] Northern Ireland Statistics and Research Agency, The Arches Centre, 11-13 Bloomfield Avenue, Belfast, Northern Ireland

Abstract

The Northern Ireland census of population is important as it provides statistical information about the population and households for all parts of the country. The information is essential for Government, business and the wider public and so substantial preparation is involved to ensure its smooth running and that it is acceptable to the public. This talk will look at recent developments in the preparation for the 2001 census.

The Government announced their proposals for the 2001 Census in a White Paper published in March this year. The White Paper covers issues such as questions to be included, confidentiality and computer security, and changes to the traditional census methodology.

It is almost certain that some people will be missed by the Census. Historically census output has been restricted to those enumerated, meaning census output appears to be superficially different from the mid-year population estimates. In 2001 it is proposed to adjust the census database to allow for non-enumeration. A key element in this will be a large scale Census Coverage Survey which will be carried out shortly after the 2001 census.

A census rehearsal will take place in April 1999, with the aim of testing the procedures for delivery and collection of the census forms, and the systems for processing the data and producing outputs.

When 2001 Census output becomes available, users will inevitably compare the results with the annual mid-year population estimates. Work is underway to quantify the accuracy of the mid-year estimates series.

The talk will focus on some of the more statistical issues regarding the census.

Statistical Inference, Likelihood and its Variants

Denis Conniffe[1]

[1] The Economic and Social Research Institute, 4 Burlington Road, Dublin 4, Ireland. E-mail: Denis.Conniffe@esri.ie

Abstract

Likelihood means different things to different people. For some, Edwards (1972) being a prime example, the Likelihood Principle - that all the information the data provides concerning the relative merit of two hypotheses is contained in the ratio of likelihoods - is the starting point that can be expanded into a whole system of inference. For Bayesians, the Likelihood Principle is a natural consequence of their own starting points. But for theoretical frequentists, it is an irritation that cannot be binding. As for the vast majority of ..metricians (econ, psyco, bio, etc), they are happily unaware of the Principle and take likelihood to mean techniques in estimation (Maximum Likelihood) and hypothesis testing (Likelihood Ratio tests, etc.) They would probably justify the techniques on the basis of wide applicability and good properties.

In fact, though ..metricians have not been slow to innovate in achieving applicability, while retaining the word 'Likelihood'. From early on, econometricians got rid of some unwanted parameters by working with 'Concentrated Likelihood', later called 'Profile Likelihood' in mainline statistics. As for good properties, biometricians from Student onwards have been unhappy with ML estimators of variance components. Some quietly corrected for bias, or even used other estimators, but with the arrival of Residual ML (Patterson and Thompson, 1971) they could return to the likelihood fold, even if unclear what REML is. Sometimes, it amounts to factorisations of the likelihood of the kind used by Kalbfleish and Scott (1970) to base inference on marginal or conditional likelihood. Even factoring out some parameters, other unwanteds could remain and a plethora of authors have proposed likelihood variations - Modified Profile, Conditional Profile, Canonical, Adjusted Profile, Approximate Conditional Profile, etc. Some of these estimators had near relatives, or perhaps previous existences, under other names, with different authors and with different, perhaps sounder, justifications. But a 'Likelihood' in the title brought an estimator under an

approved umbrella. The downright absurdity of ML in some non-regular cases has led to perhaps more extreme variations such as the Maximum Product of Spacings (Cheng and Amin, 1983; Ranneby, 1984), although in spite of efforts (Cheng and Illes, 1987), they do not seem to have been allowed under the brolly.

As estimators have increased, optimality claims have advanced. Given some regularity conditions, maximum likelihood estimators are asymptotically (1st order) efficient, as are many other estimators. But higher order efficiency for ML is claimed in the texts, although the small print reveals the estimators have been adjusted to remove bias. They are not actually ML estimators, although they may be one of the variants. The effect is to obscure the fact that ML may be inferior to other procedures. That might question the value of computational algorithms that iterate complicated likelihoods to ML solutions.

Similar processes have occurred with hypothesis testing. The 'likelihood based' testing methodologies - originally Score, LR and Wald - have expanded to include many members such as Modified Score, Conditional LR and adjustments to them. Mukerjee and Chandra (1991) is a good example of the development. As with estimation, the meaning of terms has expanded beyond their initial content. There are situations (for example, when a nuisance parameter is unidentified under the null) where a LR test does not exist, but a modification does, and gets rechristened the LR test. Again, this may disguise the true situation.

There are other starting points besides likelihood, which provide wider families of estimators and test criteria and sometimes include the likelihood based methods as special cases. You could start from Estimating Functions (for example, Godambe, 1991), which has appealed to some, or from Distance Measures (for example, Burbea and Rao, 1982) between probability distributions. Taking relative entropy or Kullback-Leibler information, (Shannon, 1948; Kullback, 1959) as the measure connects to likelihood, since entropy is proportional to minus the expectation of the log likelihood. I have argued (Conniffe, 1987, 1990) that maximisation of the expected log likelihood, and the consequent zero expectation of the score vector, is the 'true' key to estimation and testing in the regular case and in some nonregular cases (Conniffe, 1999). Taking other distance measures can sometimes make sense though (Ullah, 1996). In general, Information Theory is almost a parallel universe to Statistics with some corresponding techniques (for example, the Minimum Message Length criterion).

This talk is fully in the frequentist framework, but it will follow some paths that may not be familiar to all. I think they do get further than apparently well signposted highways. Even if you do not agree with this, the going will

be mathematically easy with interesting oddities along the way.

References

Burbea, J and C.R Rao (1982). Entropy Differential Metric, Distance and and Divergence Measures In Probability Spaces: A Unified Approach. *Journal of Multivariate Analysis* 12, 576-579.

Cheng, R.C.H and N.A.K Amin (1983). Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin *Journal of the Royal Statistical Society B* 45, 394-403.

Cheng, R.C.H and T.C.Illes (1987). Corrected Maximum Likelihood in Non-Regular Problems. *Journal of the Royal Statistical Society B* 49, 95-101.

Conniffe, D (1987) Expected Maximum Log Likelihood Estimation. *The Statistician* 36, 317-329

Conniffe, D (1990) Testing Hypotheses with Estimated Scores. *Biometrika* 77, 97-106

Conniffe, D (1999) Score Tests When A Nuisance Parameter Is Unidentified Under The Null Hypothesis. *Journal of Statistical Planning and Inference* (In Press)

Edwards, A.W.F (1972) Likelihood. *Cambridge, CUP*

Godambe, V.P (1991) Estimating Functions. *Oxford, Clarendon Press*

Kalbfleisch, J.D and D.A.Sprott (1970) Applications of Likelihood Methods To Models Involving Large Numbers of Parameters. *Journal of the Royal Statistical Society B* 32, 175-194

Kullback, S (1959) Information Theory and Statistics. *Wiley, New York*

Mukerjee, R and T.K. Chandra (1991) Bartlett-Type Adjustment For The Conditional Likelihood Ratio Statistic of Cox and Reid *Biometrika* 78, 365-372

Patterson, H.D and Thompson, R (1971) Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika* 58, 545-554

Rannerby, B (1984) The Maximum Spacing Method: An Estimating Method Related To The Maximum Likelihood Method. *Scandinvaian Journal of Statistics* 11, 93-112

Shannon, C.E (1948) A Mathematical Theory of Communication. *Bell Systems Technology Journal* 27, 379-423 and 623-656

Ullah, A (1996) Entropy, Divergence and Distance Measures With Econometric Applications. *Journal of Econometrics* 49, 137-162

Bayes' Theorem is Not a Case of Life or Death. It is Much More Important Than That.

D. Sprevak[1] and D. Davison[2]

[1] Department of Mathematics, University of Ulster and [2] Department of Applied Mathematics and Theoretical Physics, The Queen's University of Belfast

Abstract

Recently a young boy was lost in the Arizona desert. The rescue services mounted a set of searches. The assessment of the probabilities that the boy was in different areas was modified after each unsuccessful search using Bayes' Theorem which incorporates information gained by preliminary searches. The boy was found unharmed.

This paper illustrates the procedure for upgrading the prbabilities of areas and it shows that the probability that the missing person is found in the next search is maximised when the best teams are assigned to the most likely areas.

Can You Randomly Generate Numbers?

Philip J Boland[1]

[1] Department of Statistics, National University of Ireland - Dublin, Belfield, Dublin 4, Ireland E-mail: Philip.J.Boland@ucd.ie

Abstract

The generation of random numbers is an essential statistical tool used in simulation, sampling, scientific experiments, numerical analysis, decision-making, gambling and computer programming. Normally one uses a random number table, or some pseudo-random device on a computer or calculator to generate random numbers. Most of these generators rely on linear congruential methods, but there are other alternative techniques. The challenge of creating good random number generators were laid down by John Von Neuman in 1951 when he said "Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin." Is it conceivable that individuals have the capacity of being good generators of truly random numbers? In particular if **you** were personally asked to randomly generate a stream of random digits or numbers, do you think you could do a respectable job? Most Statisticians would believe this is very unlikely due to various hidden biases that individuals must seemingly possess. But what types of bias might we expect to observe when people attempt such a simple but delicate task? For example are they likely to select some numbers more frequently than others, or spread out numbers more than one might naturally expect from a truly random generator? In order to investigate these and other possible forms of bias, an experiment was conducted where individuals were asked to perform as random number generators. A large number of first year university students in statistics and mathematics were asked to "randomly" generate a sequence of 25 digits from 0,1,2,3,4,5,6,7,8,9. An analysis of the results gives some interesting insights into how the human mind seems to over-compensate in an attempt to be fair and balanced, consequently showing certain biases and in particular the tendency to avoid clustering and repetition when selecting digits.

Reconstruction of Motion Blurred Images

Kingshuk Roy Choudhury[1]

[1] Statistics Department, UCC, Cork, Ireland E-mail: kingshuk@stat.ucc.ie

Abstract

Motion blur occurs when a moving object is 'photographed' by an imaging device with a relatively slow 'shutter speed'. An example where such a problem might occur is when the police wish to decipher the registration number of a getaway car from a snapshot. Other applications occur in fields such as astronomy, remote sensing and medical imaging. If the original object was moving at a constant known velocity and the blurred image is known without error, it is possible to use a reconstruction process to recover the original image without any loss of accuracy. However, in practical situations the blurred image is contaminated by noise artifacts due to the nature of the imaging process. However small these artifacts maybe, their effect is magnified in the reconstructed image. This error magnification is a characteristic of a large class of problems known as 'inverse' or 'deconvolution' problems. Reconstruction accuracy can be improved if we incorporate known constraints on the original image, such as positivity. However, such reconstructions are difficult to efficiently implement and analyse. Another factor that can complicate reconstruction is when the velocity of motion is unknown and/or non-constant.

Slepian ('68) proposed a least-squares reconstruction method for the motion blur problem. This method can be efficiently implemented using the Fast Fourier Transform (FFT) and performs reasonably when the noise levels in the blurred image are low. After a hiatus, Vardi and Lee ('93) proposed an algorithm based on a statistical model for the data. They suggested that this algorithm improves upon the FFT based algorithm in terms of reconstruction accuracy. Roy Choudhury and O'Sullivan ('98) did a systematic comparison of these two methods for one-dimensional signals which justifies this claim (Fig 1 shows why). Additionally, it was demonstrated that a constrained least squares based algorithm would perform comparably to the EM algorithm. Spurred by a general growth of interest in the technique

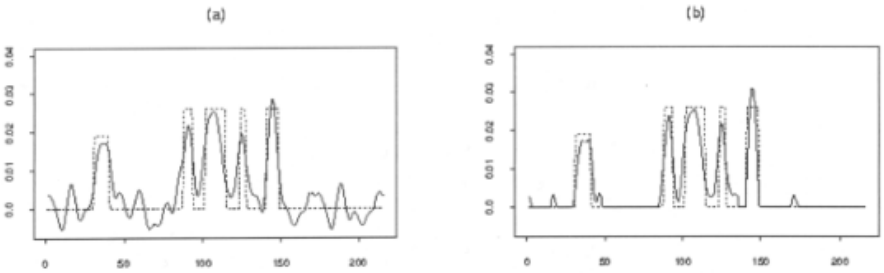

FIGURE 1. (a) Least Squares vs. (b) EM based reconstruction.

of 'interior-point' programming, recent work by Groeneboom ('96) and others has shown that in related situations algorithms based on interior point methods are computationally more efficient than EM.

In Figure 1 above, the true signal is shown in dotted lines and the reconstruction in solid lines on both plots. It can be seen that significant negativity artifacts appear in the unconstrained reconstruction. Both reconstructions were done on the same dataset and were spearately optimally regularised to give minimum Integrated Squared Error (Choudhury and O'Sullivan '98). Part of this work is joint with Piet Groeneboom, Dept of SSOR, Tu Delft and Finbarr O'Sullivan, Statistics Department, UCC.

Theoretically quantifying the accuracy of constrained reconstruction methods such as the EM or interior point based methods is not easy. In fact an exact computation of this nature is infeasible except for very simple images. Even for an asymptotic computation, the difficulty lies in the fact that for constrained methods, the reconstructed image is not a differentiable functional of the original data, whereas almost all standard asymptotic analyses in statistics are based on a differentiability argument. Roy Choudhury and Goreneboom (1999) have suggested a method of analysis for the EM based estimator based upon the Kuhn-Tucker like optimality conditions for a constrained problem. At present, this approach only works for one-dimesional signals, whereas in practical situations, one is interested in two or higher dimensional images. Extension of the existing work to these images will be a useful future contribution. A similar theory for the interior point based estimator also needs to be developed.

References

Dobrushin R, Groeneboom P, Ledoux M (1996). Lectures on Probability Theory and Statistics *Springer-Verlag, Berlin, New York*

Roy Choudhury K and O'Sullivan F (1998). An Analysis of the Role of Positive and Mixture Model Constraints in Poisson Deconvolution Problems *Journal of American Statistical Association*

Roy Choudhury K and Groeneboom P . Asymptotic Analysis of the Maximum Likelihood Estimator for the Motion Blur Problem *Manuscript in preparation*

Slepian D (1967) Restoration of Photographs Blurred by Image Motion *Bell Systems Technical Journal, Dec*

Vardi Y and Lee D (1993) From Image Deblurring to Optimal Investments: Maximum Likelihood Solutions for Positive Linear Inverse Problems *JRSS B, 55: 569-612*

Reliability Growth Modelling

John Donovan[1] and Eamonn Murphy[2]

[1] Institute of Technology, Sligo and [2] University of Limerick

Abstract

Reliability growth monitoring is an important tool for monitoring and tracking the reliability improvement achieved during product development. Many models have been proposed and are in use, although Duane's model remains the primary graphical model for electronic systems. As failures occur during a development test programme, their times to failure are recorded. The Duane model then represents the relationship between the cumulative Mean Time Between Failures and the cumulative test hours. Duane's model plots as a straight line when plotted on log-log paper.

There are a number of inherent limitations associated with Duane's model, one being that early failures have a high influence on both the graphical display and the slope of the line. Therefore should one attempt to utilise it as a means of observing growth during testing, the resulting graph is overly affected by those failures occurring early in time. Secondly, if a number of failures occur towards the latter part of the test, these points tend to be clustered together due to the nature of the 1n (Cumulative Time).

This paper presents a new reliability growth model derived from variance stabilisation transformation theory. The new model is simpler to plot and fits the data more closely than the Duane model over the range of Duane slopes typically observed during reliability growth programmes for electronic equipment. The problems inherent in the Duane model of providing too much influence to early failures is overcome.

Over 6,000 computerised simulations of reliability growth data were performed and the results indicate that for Duane slopes less than 0.5, the new model is more effective. The leverage and influence aspects of both models are evaluated by means of Cook's distance and

supports the assertion that the Duane model is unduly influenced by the early failures. The new model, on the other hand, is influenced by the latter failures which should be more important than the earliest failure. It is unrealistic to believe that after possibly 100,000 hours of testing, the earliest failure should remain the most influential.

Both models are shown to be mathematically equivalent in their capability to fit the observed data when the Duane slope is 0.5. For slopes above this, the Duane model is more effective while below this, the new model more accurately fits the data. Typical reliability growth programmes conducted on electronic equipment report Duane slopes of less than 0.5, implying that the new model is the most suitable for such applications.

Reliability growth models are used predict the total test time required to achieve a particular reliability goal, referred to as the instantaneous Mean Time Between Failures. The new model has a further advantage over the Duane model as it leads to reduced test times for achieving the specified reliability goal.

The applications of the new reliability model can be extended to software development programmes where reliability growth modelling techniques are also employed.

Finally two published datasets from the development projects of both hardware and software products are presented and these confirm the simulation results that the new model is more appropriate than the Duane model.

Do Oral Contraceptives Affect The Risk Of Breast Cancer? A Clustering Approach To Meta-Analysis

Vicki Livingstone[1]

[1] Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland. E-mail: victoria.livingstone@ul.ie

Abstract

When combining results from separate studies in a "fixed effect" meta-analysis, it is assumed that all the studies come from the same population and test the same hypothesis. In reality, this is rarely the case since variability often exists between study designs, between the countries where the studies were performed or the way in which the information was obtained.

Cluster analysis is designed to create homogenous groups. We propose an approach based on K-means clustering that partitions the studies into groups such that studies within a group are similar and studies in different groups are dissimilar. A separate "fixed-effect" meta-analysis is then performed on each group. Likelihood-based methods are discussed for selecting the optimal number of groups and, in particular, for testing the hypothesis that the studies constitute one homogenous group.

We illustrate this method using data gathered by The Collaborative Group on Hormonal Factors in Breast Cancer. This group has brought together and reanalysed the world-wide epidemiological evidence on the relation between breast cancer risk and use of hormonal contraceptives.

References

Collaborative Group on Hormonal Factors on Breast Cancer (1996), Breast Cancer and Hormonal Contraceptives. *Lancet* 1713-1727 No 347.

Practical Issues Around the Data Mining Process

Martin Duffy[1]

[1] SAS Institute, 125 Lower Baggot Street, Dublin 2, Ireland

Abstract

Data mining is becoming one of the most talked about topics in Information Technology today. There are many stories about how organisations have saved vast amounts of money by successfully implementing data mining, the problem is how to go about this process.

1 Introduction

Data Mining is a process of finding valuable information in the vast quantities of data that make up corporate computing. It relies on techniques that come from areas such as graphics, statistics, OLAP and artificial intelligence. By allowing users to find relationships, outliers and patterns, data mining turns these data into meaningful business information. But is it all as simple as it sounds? All you have to do is get a big black box and throw all the data you can at it, then all your business problems will be solved in one big bang. This approach is sure to succeed: succeed in generating a large bill for your organisation and little else.

Data analysis is only one stage in the whole process which makes up data mining. To ensure the success of this prospecting exercise we have to take a holistic approach. Data mining is not just a technology, it is a complete business process and as such has to be supported by many parts of the business for the benefits to be reaped. This is the basis of end to end data mining.

2 Why an end to end solution is important

Data mining requires several components to make it successful. Just as gold miners spend time surveying the land to ensure the right geo-

logical structure to yield gold, the data miner has to do a similar task. Data miners must first decide what they are looking for and then determine what data would be required for this. To achieve this, first of all the business questions to be asked have to be clearly defined, as must an achievable goal. It is no use defining a goal of 100% response to a mailing when your current hit rate in 2%. A goal that defines the change in profitability to the organisation is more useful, e.g. to cut the cost of mailing by 25%. This type of goal is achievable by reduce the number of people mailed and/or increase the number of respondents.

In order to decide what data would be required, a team of data analysts, business analysts and IT personnel must be formed. The roles of these people are to define the business drivers, the data that is available, data that can be sourced from other places, ways of capturing new data and types of analysis to be used to answer the questions. If this task is not undertaken first, the project may progress to an advanced stage before finding that you were digging in the wrong place.

Once the business questions and data have been defined, the next stage is to warehouse the data. This provides the raw material for the data mining exercise. The better the data the more likely it is that value will be extracted from it. Exactly how to go about this is discussed in the next section.

It is only now that the data analysis process can begin. The methodology for this stage, that SAS Institute would prescribe, is called SEMMA. This methodology is the steps that need to be taken in order to analyse the data in a structured fashion. SEMMA stands for :

- Sample
- Explore
- Manipulate
- Model
- Assess

Each one of these steps is visited several times in order to ensure that the final model is the most correct. Thus SEMMA is an iterative and interactive process for data analysis. At any stage it may be necessary to revisit the definition or warehousing stages, as the data analysis stage yields more information about the data.

When a suitable model has been chosen, the next stage is to present the results to the enduser, usually someone who has little or no data analysis knowledge, in a way that can be clearly understood. For this reason an easy to use meaningful reporting system is a must, such a system could be a paper report or more likely today is the use of an EIS system.

All these stages so far have not generated any return on investment, in fact they have cost money. It is only when the business takes actions that the investment is realised. In order to assess the return on investment given by data mining, it is necessary to first report on the current position. It is true of most organisations who undertake data mining that they do not know their current position. Therefore they cannot assess the gains from data mining. It is only with after such reports that the benefits from any actions taken can be assessed. Thus the goals for data mining should not only be achievable but also actionable.

3 Role of the data warehouse

The role of the data warehouse in data mining is to simply provide the data. The process of data warehousing adds value to the data in several ways that support the data mining function. The Meta Group "recognise that the data warehouse model provides the basic infrastructure for data mining". Data warehousing does this by providing clean, consistent, time variant, stable data that is the fuel for any successful data mining engine.

The quantities of data used in data mining can vary widely. While it is usual to develop first stage models on samples of data, usually thousands of records, the final model is usually run against the entire population, this may be several millions of records. While the data warehouse may hold all of this data, it is common for at least some of it to have been summarised. In very general terms summarised data is not ideal for data mining as it has already had some patterns removed.

Without data warehousing the process of data mining becomes much more expensive, time consuming and less attractive to the business. By having a data warehouse that can be used by several business functions, the risk associated with undertaking data mining is significantly reduced. This means that benefits can be seen across the business in a variety of projects from data mining to reporting and financial control.

4 Integrating Data Mining and Data Warehousing

When the data warehouse is being built the purpose of the warehouse has to be taken into account. A data warehouse which has been solely designed for reporting may not be structured appropriately for data mining. One of the problems that often arises is how to handle missing values. In reporting applications it is usual to either ignore, or to replace missing values with a zero. In data mining the question of why the data is missing is more important. For example, if a customer does not give their age, that is information about the type of person that customer is. In order to support the data mining function, an existing data warehouse may require some modification to its cleaning and transformation engines.

Transformation of existing data in the warehouse may also have to take place. One very common occurrence of this is when the data on gender is held in character format. While this is idea for reporting, a numeric format is more desirable for data mining. This can lead to duplication of data within the warehouse which is the very thing that data warehouses were designed to stop. In order to avoid this problem two approaches can be used. The first is to store such data in numeric format, and display it as full text strings. However not all software supports this functionality. The second approach is to build a data mart (a small data warehouse) to be used for the purpose of data mining.

The data architecture which underpins many data warehouse is the star schema. The star schema is so called because its diagram looks like a star consisting of a central "fact" table (storing measures) linked via keys to several "dimension" tables (storing attributes and classifications). For example, a customer table might also hold location information. This data might be useful for data mining purposes, but is not stored in an easy to access way for data mining. While data mining can work with the star schema, this data architecture does require that tables be joined. The problems really start with the snowflake schema, which goes even further in having branching tables. This requires many joins and can be resource intensive. This can significantly slow the load time of the data for data mining applications.

While data warehousing is currently being driven by the reporting needs of today's businesses, and then expanded out to include data mining this approach is far from ideal. More often than not the data warehouse is designed and populated for reporting, and as discussed

above this can cause problems. However, the main effect is even more worrying. Data that is not required for reporting purposes is often not captured until required. While this is perfectly feasible for reporting, which is usually looking at snapshots of the business at a given point in time, it is not acceptable for data mining.

The problem with data mining is that it requires historical data which may need to go back at least several months. If this data is not readily accessible, the data analysis phase of the project may need to be put on hold until the data is available. This loss of time may lead to a loss of competitive edge. Thus the data requirements for a data mining project should be completed as early as possible. This ensures that when the data analysis part of the project starts it has the raw materials required for success. How much data is required by the data analysis is critical. If you do not have enough data you may get misleading results or not be able to use the best analysis method. A rather simple example of this is if we look at a company selling ice cream. If this organisation only has 6 months of data available, let us say Oct-March, then the results of a data mining exercise will not give any valuable information about summer trends.

5 Importance of end user reporting

The whole point of data mining is to develop new information to give your organisation a competitive edge. This can only be successful if the business community can understand and utilise the information. To this end the reporting of the results from data mining is essential. It is no use building the best model to predict customer behaviour if nobody in the organisation can understand it. If people cannot understand the results of the model, then the model cannot add to the understanding of the customer. Thus providing timely, accessible, understandable reporting to the business is a necessity to gain the benefits from data mining.

The majority of business users want to see a report rather than the test results of the accuracy of a data analysis technique. The report may be based on the output from the data analysis method, but must be meaningful in a business context. One of the most common ways to provide this useful information is in the form of paper reports. In recent years this has changed to be more computer based, with EIS and OLAP systems leading the way.

EIS and OLAP systems have several advantages over paper based systems. Some of the more advanced EIS and OLAP systems can run a predictive model when the user enters that particular report. This means that the report that is displayed is the most up to date possible. Thus the traditional 3 month wait for the end of month forecast can become a thing of the past. A more usual method, however, is to have a results file which is regularly updated. This means that the predictive model can be run outside of peak office hours allowing for quicker display of the results report.

With enduser reporting there is also another advantage that EIS/OLAP systems can provide. As they are end user tools they can be used for simple data mining by the business community. This means that the end user can take the initiative in finding anomalies or patterns. This can then feed more advanced data analysis methods to add value to the data under examination.

6 Who owns what?

As discussed earlier, data mining is a process that involves many different disciplines within an organisation. So the question of "who owns what?" must be asked. The simple answer to this is that the success of any data mining project relies on everybody involved taking responsibility for all aspects of the process. In practice, it is a little more complex than that.

As discussed earlier, a team to define the data sources required is essential. Once this team has defined the data the next question is how to warehouse the data. While data warehousing is primarily an IT task, both the end users and data analysts, i.e. those who understand the data and who understand how the data will be used, must have major input. If the data warehouse is ring fenced by IT, the requirements of the usage of the data may not be fully understood. When this happens the project can give either unreliable results or be delayed significantly.

The data analysis stage is usually the prime responsibility of the data analyst, however IT must ensure that the data warehouse is accessible, and end users may be required to help assess the validity of the model. Here we make an important distinction between the end user and the data analyst. Some data mining tools are designed to be extremely easy to use, and therefore people who have little or no understanding of what can be done with data think they can use them

safely. This can be dangerous, so it is important for the analysts to undertake the data analysis. Where easy to use tools are applicable for endusers is in applying the models built for them. Some of these tools even allow administrators to lock parts of the data mining flow. This means that only qualified uses can change the settings.

Many of the analytical techniques used in data mining are very difficult for untrained people to understand, and even harder to use correctly. It is usually best for these methods to be dealt with by experts in the analytical side of data mining. The results from these techniques can have a major impact on the business and thus should be used by experts in that area, that is the end users.

Once the data analysis has been completed, the end users and IT have to work very closely together to ensure the end user reporting phase is successful. End users must ensure that the format and content of the reports are meaningful, and IT must provide the mechanism to produce these reports. The data analyst must also supply either the model or the results of the model to drive the reporting process. Of course the actions to be taken, in light of the new information that data mining has provided, is the responsibility of the entire business.

So the team of people who first define the data needs must work together throughout the lifetime of the project, and all have tasks and responsibilities at each stage. It is only with this team based approach that the potential of data mining can be realised.

7 Keys to successful data mining

Data mining is not just a set of data analysis techniques. Rather it is a process of organising data and extracting information which has to be delivered in a meaningful, accessible and timely way. It may involve a change in attitude to data: from being a liability that is expensive to store, into one of the most valuable assets that your organisation holds. In order to calculate the return on investment from data mining, it is first necessary to develop reports on your current state, thus any improvements can be measured.

In order to make this transformation, from data to information, a data warehouse provides the required infrastructure. Data mining must be designed into the data warehouse and the correct data must be identified and collected. This should be done as early as possible.

One of the essential differences between data mining and reporting is that data mining requires large amounts of historical data.

If data mining is being undertaken on an existing data warehouse, there may be a requirement to change the underlying data in the warehouse as well as the cleaning, loading and transformation mechanisms. The data may also not be stored in the best layout for the data analysis phase to take place, thus the data repository may also need to be looked at.

SAS Institute recommends the SEMMA approach to the data analysis phase. This should not be seen as a linear approach, rather an iterative and interactive way to analyse data. It is also likely that changes will have to be made to both the sourcing of data, and the data warehouse, in light of the experiences of the of the data analysis phase.

Any model is only useful if it is used. Therefore reporting on the results of the data analysis phase must be in a form that the business can understand. It is no use filling reports with statistical jargon when it is the effect on the bottom line that counts. To this end, the role of the analyst should be to ensure the quality of the final model, the role of the end user should be to assess the information for the business. End users who are not sufficiently trained should not build models - they will get results but cannot guarantee the quality of the results.

A team based approach to data mining allows each department to use their resources in the best way while the business gets the best results. This team should include end users from the business, data analysts and IT department. It is very important that this team owns the entire process and all members have a role to play at each stage.

The final goal of data mining should be a realistic, measurable, business driven and actionable result. It is only when the results from the data mining process are actioned that the process adds value to the business. If data mining were a simple "put any data in and get right answers out" process it would be old hat by now. The keys to success are planning, teamwork and having business driven actionable goals.

Estimation with Poisson sampling

J. M. Horgan[1]

[1] Dublin City University, Dublin 9, Ireland. E-mail: jhorgan@compapp.dcu.ie

Abstract

We offer a large-sample criterion for which a ratio estimator with Poisson sampling is more efficient than the Horvitz-Thompson estimator, and we obtain conditions for which its gains in efficiency are greatest relative to the customary estimators used in conjunction with unequal probability with-replacement and fixed-sample-size without-replacement selection schemes. The extent of the gains in efficiency is estimated with the assistance of a heteroscedastic superpopulation model.

Keywords: Anticipated variance; Efficiency; Hansen-Hurwitz and Horvitz-Thompson estimators; Ratio estimator; Superpopulation models.

1 Introduction

Unequal probability sampling, or probability-proportional-to-size (PPS) selection, is widely used in survey sampling as a means of improving the efficiency of estimators by using some correlated supplementary information (Thompson, 1997). Ideally, PPS sampling is done without replacement, and numerous such procedures exist; Brewer and Hanif (1983), for example, lists some fifty methods. However, none are entirely satisfactory from the practical point of view. Some of the procedures do not extend beyond a sample size of two, and for those that do the second order inclusion probabilities become more complex and impracticable as the sample size increases. An exception, Poisson sampling, was proposed by Hájek (1964). It selects units with PPS and maintains its simplicity regardless of the sample size. It does however have the disadvantage that the achieved sample size is variable, and, as a result the usual unbiased Horvitz-Thompson (1952) estimator of the population total tends to have a greater variance than it does when used in conjunction with a fixed-sample-size method. It may even be less efficient than the Hansen-Hurwitz (1943)

estimator customarily used in sampling with replacement. In this paper we investigate the ratio estimator of the population total suggested by Brewer, Early and Joyce (1972) and we obtain conditions for which it is more efficient than the customary estimators used in conjunction with unequal-probability sampling with replacement and fixed-sample-size methods of sampling without replacement.

2 Methods

Suppose a population labelled $U = 1, 2, \cdots, N$ consists of identifiable but unknown units $y_1, y_2, \cdots y_N$ from which a sample of target size n is chosen using Poisson sampling in order to estimate the $Y = \sum_{i=1}^{N} y_i$. If the achieved sample size is n_a, we define the ratio estimator $\widehat{Y}_{rat} = (n/n_a) \sum_{i=1}^{N} y_i/\pi_i$ when $n_a > 0$ and 0 otherwise, where π_i is the total probability of inclusion of the i^{th} unit. For example we may have $\pi_i = nz_i/Z$ with $Z = \sum_{i=1}^{N} z_i$ for some auxiliary variable $z_1, z_2, \cdots z_N$. We compare the efficiency of $\widehat{Y}_{rat}$ with that of the customary estimators used in conjunction with fixed-sample-size without-replacement PPS selection methods. We assume a heteroscedastic superpopulation model to assess the extent of the gains in efficiency with the anticipated variance defined by Izaki and Fuller (1982).

3 Results

Using an asymptotic variance formula for the Horvitz-Thompson estimator of the total, which is accurate to order $1/N$ for all fixed-sample-size without-replacement selection methods provided that the study variable y_i is related to the measure of size z_i by the linear model $y_i = \beta z_i + \epsilon_i$, we showed that the ratio estimator with Poisson sampling is more efficient when the sample size is not small. It is also more efficient than the Horvitz-Thompson estimator with Poisson sampling when the sampling fraction $f < (1 + CV_Y^2)^{-1}$. The greatest gains in efficiency of the ratio estimator over the customary estimators used in PPS designs occur when the variance of the linear heteroscedastic superpopulation model $\sigma_i^2 = cz_i^g$ is large, or equivalently when g is near 2 in the interval $[1, 2]$.

References

Brewer, K.R.W., Early, L.J. and Joyce, S.F. (1972). Selecting several samples from a single population. *Austr. J. Statist.*, 14, 231-239.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Hájek, J. (1964). Some contribution to the theory of probability sampling. *Ann. Inst. Math. Statist.*, 36, 127-133.

Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.*, 24, 333-262.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47, 663-685.

Izaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77, No. 377, 89-97.

Thompson, M.E. (1997). *Theory of sample surveys,* London: Chapman and Hall.

Residuals and Influence in Forecasting

John Haslett[1]

[1] Department of /Statistics, Trinity College, Dublin 2, Ireland. Email: John.Haslett@tcd.ie.

Abstract

Classically, residual analysis in time series modeling is concentrated on either the one-step ahead forecast errors or - in ARIMA modelling - on estimates of the underlying random innovations to which forcast errors are closely related. It is suggested here that the most appropriate residual - for the purpose of assessing influence - is neither of these. It is rather the 'leave-one-out cross-validation residual' in which the datum being investigated is indeed contrasted with the prediction from the rest of the series, but now including not only the past but also the future.

Interval Estimation of Effective Doses When a Logistic Dose-Response Curve is Incorrectly Assumed

Y.X.Huang[1] P.Harris[1], S.P.J.Kirby[2] and J.C.Dearden[3]

[1] School of Computing and Mathematical Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK, [2] Biometrics Department, Central Research Division, Pfizer Limited, Sandwich, Kent, CT13 9NJ, UK and [3] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool, L3 3AF, UK

Abstract

A number of recent studies have looked at various methods for interval estimation of the median effective dose, or of a more extreme effective dose, for binary response data when a logistic dose-response curve is correctly assumed. Here we focus our attention upon the situation in which the assumed logistic dose-response curve is incorrect and address the robustness of six interval estimation methods for the ν% effective dose (ED_v), where ν is a pre-specified percentage, and the true dose-response curves are, respectively, either the probit model, the cubic logistic (CL) model (Morgan, 1992, p 156) or the asymmetric Aranda-Ordaz (AO model (Morgan, 1992, p 147).

Keywords: Binary data; Bootstrap; Confidence interval; Robustness.

The six interval estimation methods for the ED_v considered here are:

(i) the likelihood ratio (LR) interval (Williams, 1986);

(ii) the score test interval (Harris et al., 1999);

(iii) and (iv) the Fieller interval and the interval based upon the delta method. These intervals can be represented together as being the set of μ_{v0} satisfying

$$H(\tilde{\mu}_v) = (\hat{\beta}_o - d_v + \mu_{v0}\hat{\beta}_1)^2(\nu_{11} + 2\nu_{12}\tilde{\mu}_v + \nu_{22}\tilde{\mu}_v^2)^{-1} \leq z_{0.05}^2 \quad (1)$$

where $\tilde{\mu}_v = \hat{\mu}_v$ for the delta interval (sometimes termed the asymptotic confidence interval) and $\tilde{\mu}_v = \mu_{vo}$ for the Fieller interval, $\beta = (\beta_o, \beta_1)^T$, $\mu_v = (d_v - \beta_o)/\beta_1, d_v = 1n[\nu/(100 = \nu], V = (\nu_{ij})$ is the estimated asymptotic variance matrix of $\tilde{\beta}$ and $z^2_{0.05}$ is the upper 0.05 point of the χ^2 distribution with one degree of freedom. In all cases the addition of a circumflex denotes the maximum likelihood estimator (MLE);

(v) the λ_{1-} intervals for μ_v (Huang et. al, 1998) are given by the set of μ_{vo} satisfying (1) in which $\tilde{\mu}_v$ is represented as $\tilde{\mu}_v = \lambda_1\mu_{vo} + (1 - \lambda_1)\tilde{\mu}_v$, where $0 \leq \lambda_1 \leq 1$;

(vi) the λ_2 intervals for μ_v (Huang et. al, 1998). These are the set of μ_{vo} satisfying

$$\lambda_2 H(\mu_{vo}) + (1 - \lambda_2)H(\tilde{\mu}_v) \leq z^2_{0.05} \text{ for } 0 \leq \lambda_2 \leq 1. \qquad (2)$$

The λ_{1-} and λ_{2-} intervals contain the delta and Fieller intervals as special cases corresponding to, repsectively, $\lambda_j = 0$ and 1 (j=1,2). The purpose of the λ_{j-} intervals is to achieve a compromise between the often-observed liberalism of the delta interval and the often-observed conservatism of the Fieller interval. In practice, for a given set of data, a value of λ_j for use in the λ_j- intervals needs to be chosen. If the assumed logistic model were viewed correct one approach would be to choose λ_j on the basis of bootstrap simulation, (Efron and Tibshirani, 1993), from the fitted logistic model. Here we consider the situation in which our assumed logistic model is not correct. We propose choosing λ_j by means of bootstrap sampling directly from the observed binomial proportions at each dose level. From our bootstrap sample we construct a λ_{j-} interval for $\tilde{\mu}_v$. We repeat this process for B bootstrap samples and then choose the value of λ_j which performs best in the sense of achieving coverage, for $\hat{\mu}_v$, closest to 95%. This value of λ_j is then used to construct a 95% confidence interval for μ_v.

Restricting our attention to the cases, $\nu = 50$ and 90, we compare the six intervals for μ_v in a simulation study in which our data are generated from one of the three true dose-response curves, but we (mistakenly) regard the data as having arisen from a logistic dose-response curve. For each of the experiments, 1000 simulated data sets were generated.

The two extended models (CL and AO) are indexed by an extra parameter (respectively γ and τ) which measures their departure from

the logistic model, which itself corresponds to the respective cases $\gamma = 0$ and $\tau = 1$. We define a particular method of interval estimation as being 'robust' for the range of values (γ and τ) considered, if the coverage probability (CP) for the extended model lies within the range

$$\Omega = \{(X, Y) : -1.5 \leq X \leq 1.2, -1.5 \leq Y \leq 1.2\} \tag{3}$$

where $X = CP - 95, Y = CP_o - CP$, and CP_o is the coverage probability given by, respectively, $\gamma = 0$ and $\tau = 1$. For the probit model, X and Y are defined in the same way with CP_o and CP being the coverage probabilities when the true dose-response is the logistic and probit models, respectively. Points in Ω correspond to our observed coverage, when fitting the incorrect model, being both close to 95% (X) and also having the change in the observed coverage that would arise if we had fitted the correct model, being 'small' (Y).

Our results suggest that the $\lambda_{1-}, \lambda_{2-}$, LR and score test intervals perform well in comparison with the delta and Fieller intervals which are less robust in the presence of model misspecification. On the basis of the results presented here, one might say that the $\lambda_{1-}, \lambda_{2-}$, LR and score test intervals could be viewed as safer methods to use for interval estimation of effective doses, especially as the true dose-response relationship is usually unknown. There is little to choose among the performance of the $\lambda_{1-}, \lambda_{2-}$, LR and score test intervals, with regard to their robustness in our simulation study.

References

Efron B. and Tibshirani R.J. (1993). An Introduction to the Bootstrap. *Chapman and Hall, London.*

Harris P., Hann M., Kirby S.P.J and Dearden J.C (1999). Interval estimation of the median effective dose for a logistic dose-response curve. *Journal of Applied Statistics* 26.

Huang Y.X., Harris P., Kirby S.P.J and John Dearden (1998). Alternative Approaches to the Interval Estimation of Effective Doses for Binary Response Data. *Proceedings of RSS International Conference, 89-90. Glasgow, UK.*

Morgan B.T.J (1992) Analysis of Quantal Response Data. *Chapman and Hall, London*

Williams D.A (1986) Interval estimation of the median lethal dose. *Biometrics* 42, 641-645.

Estimating Class Attendance Rates: A Group Project for a Course in Survey Sampling

Richard Bolstein[1]

[1] Department of Applied and Engineering Statistics, George Mason University, Fairfax, VA 22030, USA

Abstract

It is generally agreed that students of survey sampling need hands on experience with an actual survey to fully understand the concepts and complexities of implementing textbook theory into real world surveys. Unfortunately, it is not only difficult and time consuming for the instructor to design an interesting survey, but it can also be prohibitively expensive, take too much time away from class lectures, and may not be feasible to carry out in a single semester.

In the past, the author has carried out successful local and statewide pre-election polls in one-semester classes. These surveys were effective in providing students experience with important concepts such as random digit dialing, interviewer bias, eligible and ineligible numbers, and non-response, and had the additional benefit that estimates of a candidate's share of the vote could be checked against the population value after the election. However, there are many negatives to a telephone survey for the classroom. First, either the instructor or a knowledgeable assistant must train interviewers and monitor the survey. Second, a telephone data center is needed, and computer assisted interviewing software is highly desirable. It may also be desirable to purchase a random digit dial sample rather than generate one, and funds may be needed for long distance calls. Finally, experimentation with different sampling designs is limited.

Over the past three years the author has been conducting class attendance rate surveys instead. The basic idea is to obtain a listing of all course offerings at the university for the current semester together with enrollment counts, time and place locations, and the name of the

instructor. Classes are then sampled during the semester to estimate the attendance rate, defined as the ratio of attendees to enrollees. Virtually all problems associated with surveys, except those intrinsic to use of the telephone, arise in this survey. For example, students have to decide what sample design to use, what estimator to use, whether or not to rely on the instructor's count, what time during the class to take the count (because of students coming in late), and what the appropriate target population is. It can be used to compare survey designs, such as simple random, stratified, and two-stage sampling as well as to contrast ratio and regression estimators with the simple expansion estimator.

The survey is amenable to students with varying backgrounds. The author has divided students into four groups. The administrative group manages the data collection and overall timeliness of the survey. The sampling group (which consists of two-three students with good computer skills) cleans the sampling frame and generates the sample or samples of classes. The analysis group (which consists of students with the best statistical skills) analyzes the data and produces tables and graphs. Finally, the reporting group is responsible for writing a report with an executive summary and giving an oral presentation of results. Best of all, the final report is published in the school newspaper, so the students can take pride in their joint project.

In this paper, the author will present details and results of one recent survey.

Fractured Steel - A Bayesian Modelling Approach

Cathal D Walsh[1] and Simon P Wilson[1]

[1] Department of Statistics, Trinity College, Dublin 2, Ireland

Abstract

Items made of metal can fail due to fatigue cracking. This process is the deterioration in the strength of a structure due to the growth of cracks within the material. These cracks develop and grow at relatively low stress levels. The process is inevitable in many materials, although it can take a long time. It is useful to make predictions regarding the lifetime of such structures, which include aircraft and automotive components.

We discuss a hierarchical population model for the growth of a family of cracks, which gives reasonable reliability predictions for data at certain stresses. An examination of the data shows that coalescence between cracks is also an important feature of the data at higher stress. We propose a model for the rate of coalescence and demonstrate how to combine the two models.

The data we have comes from laboratory testing on specimens of Steel BS1425, carried out in the Department of Mechanical Engineering, Trinity College Dublin.

The modeling is carried out in a Bayesian Framework, and the package WinBUGS is used in some of the analysis.

Keywords: BS1425, Fatigue, Microcrack, Growth, Coalescence, Poisson Process, Bayes, BUGS.

Hierarchical Repeated Measures Modelling of a Change-Point Problem

Jabulani S Sithole[1]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: j.s.sithole@keele.ac.uk

Abstract

We show a Bayesian analysis of the data about the changes in prescribing habits of certain drugs after intervention. This intervention was to influence the prescribing of these drugs by advising an increase or decrease in the dosage. The model used is the random-effects model also called the Laird-Ware mixed model, which clearly takes care of the subject specific random effects such as the intercepts and slopes. Since we wish to detect a boost or decline in the prescription of these drugs due to the intervention several months after baseline, we fitted a model that is linear but with possibly different slopes before and after this time. We also developed a community of 'off the shelf' priors corresponding to a formal expression of sceptical and enthusiastic belief and compared these with reference priors. The results obtained in this analysis are compared with the results obtained using the same dataset by a non-Bayesian approach in MLWIN and OSWALD. The analysis was carried out in WinBUGS.

1 Introduction

The following study is about the intervention programme that took place in the general practices in the Wolverhampton and Staffordshire Health Authorities by the Medicines Management Department at Keele University. The programme ran for 12 months in the period July 1994 to September 1995. There were two groups of general practices involved in the study, those that received intervention (Intervention) and those that didn't (Control).

The aim of the programme was to try and effect a change in prescribing habits of certain drugs where change was needed. It consisted of three interventions delivering a total of nineteen drug specific messages over the course of a year.

It would appear the prescribing data show great variations in prescribing between practices. Practice level data were collected 6 and 3 months prior to the intervention, the intervention period itself, 3 months and 6 months after the intervention.

The data collected in the intervention programme can be classified as longitudinal data (Repeated Measurements). Repeated measurements arise when a measurement is taken repeatedly on each of a number of subjects over a number of times. In our case we have individual practices with data collected at different times. Data of this type require special methods for handling the correlations that are typically present among the observations on a given individual practice. A first-order autoregressive structure might be assumed.

For our purposes we were interested in finding out the following: -

- Whether or not there was a significant change in prescribing of respective drugs due to the Intervention.
- Whether there is a difference in prescribing between the control practices (Non Intervention) and Intervention practices due to the intervention.
- Whether the fact that a practice is fundholding or non-fundholding has an effect on prescribing.

To illustrate how we analysed these data, we have chosen to show the analysis for Ibuprofen.

2 Model

The model chosen for this analysis is the random-effects model

$$Y_{ij} = X_{i\alpha} + W_i\beta_i + \varepsilon_i, \tag{1}$$

Where W_i is a $s_i xq$ design matrix(q *typically less than* p), and β_i is a q x 1 vector of subject-specific random effects, usually assumed to be normally distributed with mean vector 0 and covariance matrix V.

The β_i capture any subject-specific mean effects, and also enable the model to reflect any extra-normal variability in the data accurately. Models of this type have been very popular for longitudinal data since their appearance in the paper of Laird and Ware (1982).

Since we wish to detect a possible boost in ibuprofen prescription 6 months after baseline (*intervention period*), we attempt to fit a model that is linear but with possibly different slopes before and after this time. Thus the subject-specific design matrix W_i for general practice i in (1) has j^{th} row

$$w_{ij} = (1, t_{ij}, (t_{ij} - 6)^+),$$

where $t_{ij}\epsilon\{0, 3, 6, 9, 12\}$ and $z^+ = max(z, 0)$. The three columns of W correspond to individual-level intercept, slope, and change in slope following the change-point, respectively. We account for the effect of covariates by including them in the fixed-effect design matrix X_i. Specifically, we set

$$X_i = (W_i, d_i W_i, a_i W_i), \tag{2}$$

Where d_i is a binary variable indicating whether general practice i received an intervention ($d_i = 1$) or not ($d_i = 0$), and a_i is another binary variable telling us whether the practice is fundholding ($a_i = 1$) or not ($a_i = 0$). In particular, our interest focuses on the α parameters corresponding to Intervention/Non-Intervention status, and whether they differ from 0.

References

Laird, N.M and Ware, J.H. (1982). Random Effects Models for Longitudinal Data. *Biometrics.* 38, 963-974.

Multiresolution Spatial Analysis

Mitchell Morehart[1], Fionn Murtagh[2], Jean-Luc Starck[3] and Yaxin Bi[4]

[1] Economic Research Service, USDA, 1800 M Street NW, Washington DC 20036, USA. Email: morehart@econ.ag.gov., [2]School of Computer Science, Queens University of Belfast, Belfast, BT7 1NN, Northern Ireland. Email: f.murtagh@qub.ac.uk., [3]CEA/DSM/DAPNIA, 91191 Gif-sur-Yvette cedex, France. Email: jstarck@cea.fr. and [4]School of Information and Software Engineering, Faculty of Informatics, University of Ulster, Shore Road, Newtownabbey, Co Antrim, BT37 0QB, Northern Ireland

Abstract

Geographic information systems (GIS) are increasingly used as tools for topographical applications and research. A comprehensive GIS is characterized by its capabilities in the areas of data processing, analysis, and post processing. This paper explores the use of the wavelet transform as a spatial analysis tool for modeling complex multivariate geographic relationships. The use of wavelets in spatial statistics is a relatively recent phenomenon that is rapidly developing. The appeal of wavelet methods stems from their ability to process noisy data with local structures and represent discontinuities such as jumps or peaks in a function. Several examples from agricultural date are used to illustrate the exploratory data analysis inherent in the wavelet transform. The resulting maps provide a convenient means of visually conveying tremendous amounts of information. The redundant a trous discrete wavelet transform is shown to aid enormously in feature detection and exploration in the succession of resolution views of the data. Analysis is carried out through use of the MR/1 multiresolution image and data analysis package.

References

Antoniadis, A., and Pham, D.T. (1995). Wavelet regression for random or irregular design. *Technical Report, University of Grenoble.*

Bruce, A. and Gao, H. (1996). Applied Wavelet Analysis with S-Plus. *New York: Springer.*

Chui, C. K. (1992). An Introduction to Wavelets. *New York: Academic Press.*

Daubechies, I. (1988) Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics.* 27, 1271-1283.

DeVore, R. A., and Lucier, B. J (1992) Fast wavelet techniques for near optimal processing. *In Proceedings of the IEEE Military Communications Conference* 48.3.1-48.3.7.

Donoho, D. L. (1993) Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. *Proceedings of Symosia in Applied Mathematics* 47, 173-205.

Deletion Diagnostics for Balanced Linear Mixed Models

Dominic Dillane[1]

[1] Dublin Institute of Technology, Cathal Brugha SE, Dublin 1, Ireland

Abstract

Mixed linear models are widely used, arising in many areas of application. Prudent use of such models demands that the assumptions underlying their application and analysis be rigorously verified. Mixed linear model fitting normally involves method of moments or likelihood estimation techniques. ANOVA methods are frequently used for estimating variance components particularly for balanced design where they produce unique, computationally inexpensive, and intuitive estimates. However all such estimates are sensitive to outlying and influential observations and it is essential that the data analyst conduct a diagnostic check on their model fit to identify and explore such observations.

The seminal work of Cook (1977) spawned the development and widespread use of case-deletion diagnostics for model evaluation and identification of anomalous observations. This work was developed for the model $\underline{Y} = X\underline{\beta} + \varepsilon$, where ε $(\underline{0}, \sigma^2 I)$ and such diagnostics are almost universally available on commonly used computer packages. However there is relatively little published in the literature on diagnostics for the more general linear model where $\underline{\varepsilon}$ $(\underline{0}, \sigma^2 V)$ of which mixed models are a specialisation; exceptions are Martin (1992), Christensen (1992a, b, 1993) and Haslett and Hayes (1998). Even less work has been devoted to considering the effect of re-estimating the V matrix when cases have been deleted. The main challenge in developing such diagnostics is that they are normally computationally expensive usually requiring a re-estimation of the model.
In this article, we consider the computation of fixed effects and variance components diagnostic measures for mixed models. Efficient updating formulae for deletion diagnostics are developed for the case of balanced mixed linear models.

Analysis of Multidimensional Time Series with Application to Climatology

Jian Huang[1] and Finbarr O'Sullivan[1]

[1] Department of Statistics, Univeristy College Cork, Ireland

Abstract

Analysis of climate data sets is made complicated by high spatial and temporal dimensionality. Dimension reduction techniques such as principal components or empirical orthogonal function analysis have been used in this setting. These analyses have typically been used to select spatially defined eigen vectors. We explore a methodology focusing on extracting temporally defined eigen vectors. Computation techniques are developed to make the approach feasible even with very long time series.

The methodology is illustrated by application to the real climate data. A Bootstrap is used to evaluate the variability of the estimates involved.

A Survival Analysis of the Progression to Treatment for Opiate Users.

Elaine Hand[1]

[1] Mathematics Department, National University of Ireland, Maynooth, Co Kildare, Ireland

Abstract

The Health Research Board provides an annual report, which include information on opiate users in treatment and the duration of use prior to treatment. To date no detailed study statistical analysis of this data has been performed. Dean et al1 in a 1985 study of the opiate epidemic in Dublin between 1979 and 1983 did indicate that the duration of use prior to treatment was approximately 4 years and did appear to be decreasing.

Recently the European Monitoring Centre for Drugs and drug Addiction (EMCDDA) commissioned a European pilot study of this area. Using various survival analysis methods we have conducted an in-depth Irish study of this time to first treatment and compared our results with the European report.

Following a detailed analysis of various factors including age, sex, frequency of use and method of consumption, we find considerable differences between the Irish and European experience, with mean duration of use prior to treatment significantly less is the Irish situation. This verifies the observations by Dean et al1 in 1985 and has important implications for the design and implementation of first treatment programs.

References

Dean G., O'Hare A., O'Connor A., Kelly M., Kelly G. (1985). The Opiate Epidemic in Dublin 1979-1983. *Irish Medical Journal.* 78; 4.

Comiskey C.M. (1998). Estimating the Prevalence of Opiate Drug Use in Dublin, Ireland during 1996. *The Department of Health.*

Applying Logistic Regression, Probits and Discriminant Analysis to Financial Distress Prediction

Mike Feng-Yu Lin[1] and Sally McClean[1]

[1] University of Ulster, Faculty of Informatics, University of Ulster, Cromore Road, Coleraine, BT52 1SA, Northern Ireland

Abstract

In research on financial failure prediction, the data are often multidimensional, containing a large number of possibly relevant variables. To create a model from available data it may therefore be necessary to apply data reduction techniques or to partition the dataset in order to decrease its dimension. Various approaches have been applied to distinguish ongoing companies from those that ultimately fail. The statistical methods used in this study include linear discriminant analysis, logit analysis, and a probit model.

Data were extracted from UK companies listed in the London stock market for the last 10 - 20 years. This paper describes the choice of data, statistical methods, validation method, and the prediction of appropriate quantities such as the estimated probability of failure. The final section presents the most significant variables for prediction in the financial failure problem.

1 Background

The financial distress prediction problem remains of great interest to researchers as well as creditors, auditors, stockholders, and firm managers. They all have an interest in utilising and developing a methodology or model that will allow them to monitor the financial performance of a firm. Financial distress analysis can be helpful in identifying internal problems, firm evaluation by investors, and as a tool used by auditors to assist them in their job. Creditors have

a vested interest in identifying the negative developments of their borrowers. Stockholders hold the same monetary concerns. Auditors need to determine whether or not the firm's operating ability is endangered. Firm managers and board of directors can avert the crisis in advance. For all parties, there is a need to have an objective opinion on the risk of financial distress or bankruptcy as early as possible.

2 Approach

The approach of this research is as following:

1. The theoretical part The choice of variables and development of the prediction model belong to the theoretical purpose, which is concerned with the investigation and comparison of companies with distress and without distress, finding the relationship between financial ratios (the independent variables) and distress or non-distress (the dependent variable); we then construct the prediction model.
2. The empirical part Testing the theoretical model and identifying the real status of a firm are the empirical purpose. In other words, according to the prediction model based on theoretical analysis, we use the sample data to verify the prediction model.
3. Development of understanding and modelling of financial distress.
4. Identification of the financial variables which are most important in detecting the financial distress.

The following are the various steps in the design and experimental phases of the research.

Financial statement data are first collected from the UK stock market. The companies selected are composed of distressed and non-distressed companies of the same industry and approximately the same asset size. The list of failed companies is here derived from:

1. Extel UK listed securities of Negligible value,
2. Companies in receivership or liquidation. For both failed companies and ongoing companies, the financial data is drawn from the following databases:
3. Extel (Financial Times London),
4. Datastream (PRIMARK).

Once the bankrupt firms are identified, a control sample of non-bankrupt companies is usually drawn according to the industry and size (Beaver [1966]; Deakin [1972]; Schwartz and Menon [1985]; Zavgren [1985]).

Bankruptcy as a dependent variable

Studies about bankruptcy normally use bankruptcy as the dependent variable. Most researchers who study financial distress chose bankruptcy as the surrogate for financial distress. Using bankruptcy to measure financial distress seems to provide a standard that can be objectively determined (Jones [1987]).

Ratios as independent variables

McKinley et al. [????] states that ratios are the best known and most widely used of financial analysis tools. They allow the analyst to study the relationships among various components and to compare a company's performance to that of similar enterprises. Miller believes that some ratios represent cause and some represent effect.

Most researchers (for example, Beaver [1966], Altman [1968], Ohlson [1980]) have selected financial ratios as predictor variables because of their popularity and predictive success in previous research studies. Typically financial ratios are based upon data from financial statement such as the balance sheet, income and loss statement, and the cash flow statement.

The Data Analysis Procedure:

1. We use the non-parametric Kolmogorov-Smirnov test to test the normality of financial ratios.
2. We use ANOVA to compare the characteristics of the distressed and non-distressed companies' financial ratios, and observe their trend.
3. Factor analysis is employed to extract the principal factors of distressed companies for each year to examine how these factors vary with time periods.

A logit approach is selected for the preliminary data analysis because the distributions of financial ratios often violates the normality assumption. The logit approach can help us create an early warning of financial distress. Also multiple discriminant analysis and Probits will be used for the purpose of comparison and contrast.

References

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance.* (September):589-609.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Reseach.* 4:71 - 111.

Deakin, E. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Literature.* (Spring) 167-79.

Frederick L. Jones (1987) Current techniques in bankruptcy prediction. *Journal of Accounting Literature.* Vol. 6, pp 131-164.

Lin, Zhangxi, Rasf (1989) Automating Routine Analysis of Financial Data, Expert Systems In Economics, Banking and Management. pp 69-75.

Ohlson, J (1980) Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research.* (Spring): 109-131.

Schwartz, K. and K. Menon. (1985) Auditor switches by failed firms. *The Accounting Review.* (April):248:61.

Zavgren, C. (1983) The prediction of corporate failure: the state of art. *Journal of Accounting Literature.* 2: 1-37

Dynamics of Meningococcal Meningitis in Ireland

Gloria Crispino O'Connell[1] and C. Comiskey[2]

[1] School of Science, Institute of Technology, Tallaght, Dublin 24, Ireland and [2] Mathematics Department of NUI Maynooth, Maynooth, Co. Kildare, Ireland

Abstract

The most common type of bacterial meningitis in the Western European countries is caused by Neisseria meningitidis bacteria, which develop Meningococcal meningitis.

We hve formulated an epidemiological model that enables to understand the transmission dynamics of the infection. It is a set of differential equations. We have evaluated the transmission dynamics of the model.

We will present the mathematical analysis of the model dynamics and the epidemiological impatcs of the transmission. We will illustrate the results of the computer simulations and show how the infection has spread for the last 20 years. Finally we will give our first predictions on how it will spread in the next decades.

DOE in an SME - The Answer to Achieving Quality Leadership?

Christine Simms[1] and John Garvin[1]

[1] School of Management, University of Ulster, Shore Road, Newtownabbey, Co. Antrim, BT37 0QB

Abstract

The widespread adoption of Design of Experiments in industry has been attributed to the work of Taguchi in particular. However, application has been mostly confined to large organisations able to devote the necessary specialist resources to explore and develop these techniques, while SMEs (small and medium enterprises) have been slower to invest in these advanced methodologies. For SMEs operating in specialist niche markets quality leadership is vital, and for economies in which such enterprises dominate a greater awareness and use of such techniques is becoming essential to compete globally. This project focuses on the introduction of Design of Experiment techniques at Valpar Industrial Ltd, to assess their suitability for SME environments.

1 Background

Valpar Industrial Ltd. is a small privately owned company with some 50 employees located in Bangor (Co. Down), which manufactures piping for the licensed and soft drinks trade. Its speciality is Python, a flexible snake-like product which encases colour-coded tubes in a flexible insulated sleeve, enabling up to 24 different chilled beverages to be piped from keg or barrel to bar or counter, with which they have achieved a 55% share of the world market. To maintain this dominant position requires quality leadership, and the application of Taguchi methods is the preferred method of improving product quality and factory productivity, leading to the current project in conjunction with the University of Ulster.

2 The Project Objectives

The objectives of the project were defined to be:

1. To design and conduct suitable experiments which would determine and quantify the effect of various controllable variables on tube quality.
2. By improving parameter design to achieve enhanced process capability and reduced non-conformance.
3. To improve machine and material utilisation via 1 and 2.

For the initial investigation 0.5in. (external diameter) single walled high density polythene tubing, representing about 20% of total extruder output per week, was chosen. The first question to be answered was "What constitutes quality in this product?" The features identified as important to the customer were **internal** and **external,** diameter and **roundness,** along with **consistency of wall thickness** to ensure uniform strength.

3 The Initial Study

Samples were taken at random from normal production using standard settings, measured, and analysed. The results showed that wall thickness, internal, and external diameters were all within specification, and that ovality was the biggest problem.

4 Experiment 1

The first experiment was designed as an L8 orthogonal array, and investigated five factors at two extreme values, along with the interactions between two of them. This was intended to distinguish between factors which exerted considerable influence on the quality characteristics, and those whose influence was minimal. In reality it was found that all the factors investigated were significant in their effect on at least one aspect of quality.

5 Experiment 2

In this experiment an L9 array was employed, permitting the investigation of four factors at three different levels. The three most significant factors from experiment 1 were used, along with an additional factor also regarded as important. This experiment revealed the complex nature of the process and the relationships within it.

6 Experiment 3

This investigation uses an L9 array to investigate values in the immediate vicinity of the parameter settings deemed to be best from the previous expt., and thus represents a further step towards the optimum solution.

7 Experiment 4

To fully complete the picture a final experiment will be conducted to confirm the settings determined above, and to "fine-tune" the most significant factors.

8 Conclusions

The application of Taguchi methods and Design of Experiments to the problems encountered here have shown that there is considerable benefit to be gained by SMEs investing in these techniques. For processes where complex relationships exist between the various factors exerting most influence on the quality characteristics, there is no other method of establishing these relationships which is as quick and effective. This makes the technique very suitable for use within SMEs, however a substantial level of help and support from outside the organisation is likely to be required, at least in the early stages, to accelerate the learning process.