Contents

The Genetic Epidemiology of Breast Cancer

Julian Peto
London School of Hygiene and Tropical Medicine

Abstract

TEXT

Text

Estimating Risk Profile Over Time from Stored Blood Samples and Exposure Records of a Case Series

M. Reilly, E. Lawlor
Department of Epidemiology & Public Health, University College Cork, Ireland
Irish Blood Transfusion Service and Trinity College Dublin

Abstract

In some medical investigations, data might be readily available only from cases, so that standard case–control methods of estimating risk are not applicable. Such a situation arose in Ireland in 2000, when a State Tribunal of Inquiry was set up to investigate the circumstances surrounding the infection of haemophiliacs with HIV. In order to estimate the time of infection, a number of stored samples had been retrospectively tested in 1986. The question arose as to whether the collection of tests available on these patients could be used to estimate the risk of infection over time.

Using summary information on the yearly exposure to (i.e. risk from) blood products for each of these patients, end–of–year estimates of risk and relative risk were obtained from a Poisson regression model. A more refined profile of risk over time for each subjest was obtained using the detailed records of the dates and volumes of all exposures. In a final analysis, identifiers of batches of blood product can be used to apportion the risk.

The use of statistical models in this setting has two important consequences. Firstly, it enables estimates of risk to be obtained from case–cohort data, where the usual case–control methods are inappropriate. Secondly, the application of such models as data accrues, especially for newly emerging infectious diseases, can help inform policy regarding the procurement and handling of blood products in order to ensure the maximum possible safety.

Measurement of Tissue Heterogeneity based on 3-D PET Data

S. Roy, F.O'Sullivan, J.O'Sullivan, J.Eary, C. Vernon
Department of Statistics, University College Cork, Department of Radiology, University of Washington, Seattle, WA 98195, USA

Abstract

We have been developing quantitative measures of heterogeneity in the glucose consumption of human sarcoma based on 3-D FDG-PET imaging. Our goal is to evaluate the potential role of such measures for predicting treatment outcome. Previous work has found a correlation between sarcoma response and FDG uptake- however, qualitatively, more heterogeneous tumors with the same overall FDG uptake tend have poorer response. Methods: The heterogeneity measures we are developing are based on deviation of the local PET glucose utilization values from idealized models, in which the pattern of utilization within the tumor region follows a unimodal structure with simple or generalized elliptical contours. Fast and efficient computer algorithms have been developed for evaluation of these heterogeneity measures based on standard volumetric region of interest data. The heterogeneity analysis is being applied to on-going data from a clinical study in which sarcomas are imaged with FDG PET prior to surgical resection. Figure 1 shows 4 sample images together with heterogeneity values computed. The low grade tumor with high heterogeneity had poor prognosis; the high grade tumor with low heterogeneity had very favorable prognosis. To date a set of 27 such studies have been analyzed. Most (21) of theses sarcomas were of intermediate grade. The remaining six were divided equally between low and high grade. A multivariate survival analysis (using the Cox proportional hazards model) was carried out to examine how the heterogeneity measures, after adjustment for overall glucose utilization, performed as a prognostic indicator of survival. This analysis indicates that heterogeneity (even after adjustment for overall glucose utilization) has a significant (p–value is 0.021) association with survival. For tumors of the same overall glucose utilization status, the higher the heterogeneity measure the poorer the prognosis. In summary a promising method of measuring the spatial heterogeneity in the glucose utilization of human sarcomas based on FDG-PET data has been developed. This measure enhances the ability of

FDG-PET studies to predict outcome in patients with sarcoma.

Support by N.I.H./N.C.I. grant CA–65537

Analysis of Regularity in Wind Generated Wave Field Data

F. O'Sullivan, K. Roy Choudhury, G. Caulliez, V.I. Shrira
Department of Statistics, University College Cork, IRPHE, Marseille, France & Department of Mathematics, Keele University

Abstract

This talk is a follow up of the talk given in CASI 2000 on the same project. In the last talk, we described the motivation for analysing wave patterns and how the data were obtained. In the current talk, we describe progress made on the analysis of such data sets. The experiments of Collard and Caulliez [1] generated waves under various wind speed settings. These gave rise to various patterns, as shown in Fig1 (T-N). It is the object of our analysis to obtain a good representation of these patterns and thereby characterise the regularity (or otherwise) we see in these patterns.

Our method of representation of these '3–d' wave patterns is as a sum of underlying planar wave fronts. Namely, we represent the 'three-dimensional wave' $f(x, y)$, as a linear combination of components g_i, each in a different direction θi.

$$f(x, y) = \sum g_i(x \sin(\theta_i) + y \cos(\theta_i)) \quad (1)$$

Such a representation of water waves has physical justification [2]. From a statistical viewpoint, this model is similar to projection– pursuit (PP) [3], although the motivation is somewhat different. We fit the model (1) using the usual step-wise procedure as described in [3]. Usually this type of fitting is sub-optimal, but we shall demonstrate that in the case of wave patterns it does almost as well as more general procedures, such as back fitting.

The adequacy of the plane wave representation is captured in the quality of the PP fit, which can be described by a percent variance explained curve, as shown in Fig1(a). A pattern where most of the variance is explained by a few PP components would thus be more regular than one which needs more PP components to explain the same amount of variance.

However, Fig 1(a) cannot be directly used to compare the regularity of different images. This is because images of different resolution require (possibly) different amount of smoothing in the PP fitting procedure. It is well known that the smoothing parameter has a substantial effect on variability of a model fit. Therefore in devising a measure of regularity, one needs to adjust the measure for the regularity of the image. Our solution to this problem is to compare the variance explained curve of a given image to a reference curve, generated from a white noise image of the same resolution as the image in question, as in Fig1(b). Measures of regularity based on this comparison are thus independent of the value of the smoothing parameter used. Fig 1(c) shows a regularity comparison of the images shown in Fig 1 (T-N) based on this approach.

The talk will conclude with a discussion of future areas of research in this area.

References

Caulliez, G. & Collard, F., 1999, Three-dimensional evolution of wind waves from gravity capillary to short gravity range, *European Journal of Mechanics B/Fluids, 18, N3, 389-402.*

Shrira, V. I., Badulin, S. I. & Kharif C., 1996, A model of water wave "horse-shoe" patterns, *J. Fluid Mech. 318, 375-404.*

Huber, P., 1985, Projection Pursuit, *Ann. Stat. 13(2).*

Predicting Past Climates

John Haslett and Matt Whiley
Department of Statistics, Trinity College Dublin

Abstract

For some decades quantitative procedures, often referred to as transfer functions methods, have been available to make palaeo-environmental reconstructions from different types of fossil assemblages including those of pollen, diatoms, chironomids etc. Such models are calibrated on modern data, involving a wide variety of environments at some thousands of sites across Europe. Current methods do not address many of the important spatial and temporal features of such calibration data, and the models are thus inadequate for exploring detailed research hypotheses.

Fossil pollen data are available as counts of spores, for very many taxa, from radio-carbon dated slices at different depths in cores taken from lake sediments. Very many of the counts are small or zero. Our approach to the modelling of such data involves Bayesian hierarchical models. We envisage latent variables, smoothly varying in space and time, the values of which control aspects of the distribution of the observables. Counts are realizations of Poisson processes with mean values controlled by the latent variables. Observable aspects of climate, such as the mean temperatures of the warmest and coldest months, are Gaussian. All observables are conditionally independent, given the latent variables.

The solution methodology involves MCMC. The size of the data sets poses considerable technical challenges. The paper will report on some of these.

Statistical Analysis of Microphotographic Image Data in Geology

K. Roy Choudhury, F. O'Sullivan, P. De, P. Meere, K. Mulchrone
Department of Statistics, Applied Mathematics and Geology, University College Cork

Abstract

Studying the changes in rocks that take place during deformation is basic to the understanding of the tectonic history of rocks. An essential part of studying these changes is the determination of strain from microphotographic images of rock sample sections, however, the basic measurements derived from microphotographic samples are highly subjective and notoriously labor intensive. We are working to develop robust statistical image analysis techniques for analysis of microphotographic data used in determining regional strain patterns. These techniques must be capable of automatically identifying objects in the microphotograph and quantifying their associated location, size and shape characteristics. The development is being guided by a range of rock sample data made available by geologists at University College Cork and their collaborators in Sweden.

Our microphotographic device can acquire up to twenty 640X480 sub-micron resolution images per rock sample. The different images correspond to variations in optical filtering and polarization that can be used with the device. Two sample true colour images are shown in Figure 1 below. These birefringence images of the same rock sample were acquired at two different polarization angles. The visually apparent objects are grains of quartz on the order of 10 to 100 microns in diameter. Differences in the appearance of these objects at the two polarization angles reflect differences in the angles of orientation of the quart grains.

We have developed a split–and–merge segmentation algorithm for application to multiple microphotographic images of the same rock sample. This algorithm incorporates a complexity criterion that penalizes objects for lack of compactness. A single regularization parameter controls the trade–off between model fit and complexity. A segmentation of the two birefringence images is shown in Figure 1 (note that since each true colour image results in 3 grey scale images corresponding to the red, blue and green components, a total of 6 grey scale images were involved in this analysis). In qualitative terms the segmentation appears quite reasonable. Further refinement of this algorithm is needed however to incorporate a data–driven procedure for selection of the regularization parameter. Some post-processing is required to distinguish segments corresponding to quartz grains from segments corresponding to rock matrix and other structures. A possible approach to this problem based on examination of multiple polarization angles is currently being examined. A summary of results obtained will be presented in the talk.

(Supported by Enterprise Ireland grant SC/2000/104)

Figure

Modelling Persistent Correlation in Rainfall

Agus Salim
Department of Statistics, University College Cork, Ireland

Abstract

Rainfall data at hourly level show a persistent correlation that is not captured by the existing models based on the Bartlett–Lewis process (e.g. Rodriquez–Iturbe, Cox and Isham 1987, 1988, Chandler, 1997). We propose some extensions of the Bartlett–Lewis model that have a much better fit to the data. We first examine a Pareto storm inter–arrival distribution, which is partly motivated by many studies in computer science. The Pareto inter-arrival distribution has been widely used to capture the long-term correlation in web traffic data (Willinger, 1995). Since Pareto distribution is a heavy-tailed distribution it also has the potential for modelling the extreme dry run frequently encountered in rainfall data. In the proposed model, a storm inter-arrival is modelled using a Pareto distribution with scale parameter η and shape parameter ω. Each storm will trigger a Poisson process of raincells with rate β. Each raincell will have independent intensity and duration distributed according to a specified distribution. The exponential distribution is used to model the duration and intensity of raincells. After a length of time exponentially distributed with rate γ, the process of raincell generation is terminated, thus marking the end of a storm. In order to account for the effect of the multi-layer structure in the rainfall process, e.g. small mesoscale area, large mesoscale area, etc. (Waymire and Gupta, 1981), we also develop a multi-layered storm model. The structure of the model is similar to the one above except that now each arrival starts a number of storms instead of just a single storm. We fit the models to hourly rainfall data from the Valencia observatory, southwest Ireland. The original Bartlett-Lewis model and the generalised Bartlett–Lewis model with Pareto inter-arrival are compared, and both models are fitted with single and two-layered storm. A quasi-likelihood method is used to estimate the parameters of the models. From the four models, the single-layered storm model with Pareto storm inter-arrival is the best model, capturing the persistent correlation and also improving the fit of dry period distribution.

References

Boender,C.G.E.,Rinnooy, K.A.,Timmer,G.T. and Stougie,L. (1982). A stochastic method for global optimization. *Mathematical Programming, 22,125-140.*

Chandler, R.E. (1997). A spectral method for estimating parameters in rainfall models. *Bernoulli, 3,301-322.*

Cowpertwait,P. (1998). A Poisson-cluster model of rainfall: higher-order moments and extreme value. *Proceeding Royal Society London A., 454,885-898.*

Cox, D.R. and Isham,V. (1980).*Point Processes.* London: Chapman and Hall.

Csendes, T. (1988). Nonlinear parameter estimation by global optimization: Efficiency and Reliability. *Acta Cybernetica, 8,361-370.*

Diggle,P and Gratton,R. (1983). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society Series B, 46,193-227.*

Lam, W.M. and Wornell, G.W. (1995). Multiscale representation and estimation of fractal point processes. *IEEE Transactions on Signal Processing, 43,2606-2617.*

McLachlan, G.J. and Krishnan, T. (1997). *The EM algorithm and Extensions.* New-York:Wiley.

Mcullagh, P and Nelder, J.A. (1989). *Generalized linear models.* 2^{nd} Ed. London: Chapman and Hall.

Rodriquez-iturbe,I., Cox, D.R. and Isham, V. (1987). Some models for rainfall based on stochastic point processes. *Proceeding Royal Society London A., 410,269-288.*

Rodriquez-iturbe,I., Cox, D.R. and Isham, V. (1988). A point process model for rainfall: further developments. *Proceeding Royal Society London A., 417,283-298.*

Waymire, E. and Gupta,V.K. (1981). The mathematical structure of rainfall representation. A review of stochastic rainfall models. *Water Resources Research, 17,1261-1272.*

Willinger, W., Taqqu, M.S., Leland, W.E. and Wilson, D.V. (1995). Self-similarity in high-speed packet traffic: analysis and modelling of ethernet traffic measurements. *Statistical Science, 10,67-85.*

Issues in Analysing Unbalanced Data

John Connolly
University College Dublin

Abstract

Analysis of Correlated Binary Outcome Data Using Within Cluster Paired Resampling

Randall H. Rieger, Clarice R. Weinberg
Department of Mathematics, West Chester University, West Chester, PA. USA
Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC. USA

Abstract

Conditional Logistic Regression (CLR) is a commonly-used method of clustered binary outcome data analysis when interest lies in estimating the cluster-specific exposure effect, while treating the dependency arising from random cluster effects as nuisance. CLR achieves this goal by creating a conditional likelihood that does not include the baseline risks. CLR assumes that all unmeasured cluster-specific factors are aggregated into a cluster-specific baseline risk. Within Cluster Paired Resampling (WCPR) is proposed as an alternative method for analyzing clustered binary outcome data while conditioning the cluster effects out of the likelihood. WCPR allows for within-cluster dependency not solely due to baseline heterogeneity. For example, dependency may arise due to heterogenous susceptibility within each cluster.

Extensive simulation study results will be presented to compare the finite sample behavior of WCPR to that of CLR. When both CLR and WCPR are valid, our simulations suggest that the two methods perform comparably. When CLR is invalid, WCPR continues to have good operating characteristics. For illustration, we apply both WCPR and CLR to a periodontal data set that exhibits between-cluster heterogeneity in response to exposure.

WCPR promises to have a wide range of applications. In this presentation, the application of WCPR to genetic sibship-based studies will be discussed. Tests of linkage and association between a disease and either a candidate allele or marker allele can be based on sibships with at least one affected and one unaffected sibling. However, specialized techniques are required to account for within-sibship correlation if some sibships contain more than one affected or more than one unaffected sibling. Application of WCPR to such data allows testing the null hypothesis of no linkage or no association, even when sibships contain variable numbers of siblings. Two testing strategies using the WCPR procedure are proposed. Simulation results compare the WCPR testing methods to the sib transmission/disequilibrium test (S-TDT) and the sibship disequilibrium test (SDT) under various scenarios common in sibship-based genetic studies.

Some Aspects of Modelling Mean-Covariance Structures arising in Longitudinal RCTS

Gilbert MacKenzie and Jianxin Pan
Centre for Medical Statistics, Keele University, UK

Abstract

Pourahmadi (1999) provided a convenient re-parameterisation of the marginal covariance matrix arising in longitudinal studies. The new parameters, have transparent statistical interpretations, are unconstrained and may be modelled parsimoniously in terms of polynomials of time. We exploit Pourahmadi's analytical framework to model the dependence of the covariance structure on baseline covariates, time and their interaction.

In many biological and medical problems the variability of repeated over time may be influenced by the baseline covariate profile. For example, in a longitudinal study the severity of a condition may limit the evolution of response. Alternatively, the effect of intervention in a longitudinal randomised controlled trial may be to modify the covariance structure, per se, rather than, or perhaps as well as, the marginal mean. Accordingly, the assumption of a common covariance matrix in the treatment groups studied may sometimes be implausible. The rationale for our approach is predicated on the realisation that in linear mixed models the assumption of a homogeneous covariance structure with respect to the covariate space is a testable model choice.

Accordingly, we provide methods for testing this assumption. In particular, we extend Pourahmadi's results in three ways: (a) by incorporating covariates along with time into the model for the covariance structure, (b) by considering unbalanced longitudinal data and (c) by presenting additional computational algorithms which are analytically tractable. We illustrate the methods by re-analysing Kenward's (1987) cattle data and comparing our findings with Pourahamadi's analysis of the same data set.

Keywords: Covariate dependent covariance matrix; Generalised linear models; Joint mean-covariance models; Linear mixed models: Longitudinal RCTs

References

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl.Statist. :36, 296-308.*

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: *Biometrika, 86, 677-90.*

Modelling Covariance Structure in Linear Mixed Models

Jianxin Pan and Gilbert MacKenzie
Centre for Medical Statistics, Keele University, UK

Abstract

Linear mixed models (LMMs) are widely used for analysis of longitudinal data because correlation in repeated measures over time for each subject is taken into account by incorporation of random effects (e.g., Diggle, Liang and Zeger, 1994). The general form of LMMs can be written as

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i, \tag{1}$$

where, for the ith subject, Y_i is the $(m_i \times 1)$ stacked vector of m_i responses made over time, X_i is a $(m_i \times p)$ matrix of covariates, β is a $(p \times 1)$ vector of unknown fixed effects, Z_i is a $(m_i \times q)$ design matrix for the $(q \times 1)$ vector of between subjects random effects u_i, and ϵ_i is a $(m_i \times 1)$ vector of residuals, for $i = 1, .., n$ subjects. It is usual to adopt a two-stage hierarchical modelling approach in which $u_i \sim \mathrm{N}(0,\ G)$ and $\epsilon_i | u_i \sim \mathrm{N}(0,\ R_i)$ where G is the $(q \times q)$ between subjects covariance matrix, constant for all subjects, and R_i is the $(m_i \times m_i)$ conditional covariance matrix for the repeated measurements made on the ith subject, given the random effects u_i.

These arrangements lead, after integrating out u_i, to a marginal model in which $E(Y_i) = X_i\beta$ and $V(Y_i) = Z_i G Z_i^{'} + R_i = \Sigma_i$. In longitudinal data analysis, the usual parametric specification for R_i and G is assuming $R_i = \sigma_\epsilon^2 I_{m_i}$ and $G = \sigma_u^2 I_q$. In other words, the repeated measures over time are conditionally independent and the components of the random effects u_i are independent as well. In this case, the marginal covariance matrix Σ_i has the property of compound symmetry. However, the assumption of conditional independence may be unreasonable in practice since, when it holds, the correlation between measurements on the same subject is generated principally by the magnitude of the between-subject variation (Reeves and MacKenzie, 1998). Accordingly, it is usual to adopt a particular within subject covariance structure for R_i - AR(1), AR(2) and unstructured covariance models are frequently considered in the literature. Alternatively, a Gaussian stochastic process (Diggle, Liang and Zeger, 1994) may be adopted.

However, such methods are *menu-based* and problems may arise when the selected covariance structure is very different from the true structure. For example, the mis-specification may bias the estimate of the fixed effects in finite samples. Pourahmadi (1999) proposed a more flexible *data-driven* approach in which any *marginal* covariance matrix arising in longitudinal studies may be modelled using a polynomial of time.

In longitudinal studies, however, the assumption of a homogeneous marginal covariance structure is a *testable* model choice and Pan and MacKenzie (2000) generalised Pourahmadi's approach by including baseline covariates in the specification of the marginal covariance matrix arising in LMMs, i.e, from $\Sigma(t;\theta) \rightarrow \Sigma(t, \theta, x, \beta^*)$.

In the framework of LMMs, the parameters in the between subject covariance matrix G are constant across subjects and so any heterogeneity in the marginal covariance should arise as a consequence of hetereogeneity in the conditional covariance matrices R_i. In this paper, we provide a data-driven approach to detect and explain heterogeneity in the conditional covariance matrices R_i. We model the R_i parsimoniously using a regression approach and estimate the parameters by a maximum hierarchical likelihood estimation (MHLE) procedure (Lee and Nelder, 1996) which exploits the information in the estimated marginal covariance matrix. We compare our proposed procedure with standard menu selection methods, reporting results from a simulation study and the analysis of Kenward's cattle data.

References

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Clarendon Press.

Lee, Y., and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussions). *Journal of the Royal Statistical Society, Series B, 58, 619-678.*

Pan, J. and MacKenzie, G. (2000). Regression models for covariance structures in longitudinal studies. *submitted to Biometrika.*

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika 86, 677-90.*

Reeves, J., and MacKenzie, G. (1998). A bivariate regression model with serial correlation. *Journal of The Royal Statistical Society D 47, 607-615.*

Sensitivity Analysis of Longitudinal Binary Quality of Life Data with Dropout: An example using the EORTC QLQ–C30

Desmond Curran, Kristel Van Steen, Geert Molenberghs
Department of Biostatistics, ICON Clinical Research, Dublin
Center for Statistics, Limburgs Universitair Centrum, Belgium

Abstract

Introduction

Quality of Life (QoL) is rapidly becoming an integral part of clinical trials. Generally QoL is assessed using self-report questionnaires containing items (questions) with ordinal or binary response categories. Some of these items are subsequently collapsed into subscales which are also discrete in nature. In the literature these scales are frequently analyzed using the assumption of normality (possibly after transformation) or alternatively using non–parametric methods in cross-sectional analysis ignoring the longitudinal characteristics of the data. In this paper we analyze the physical functioning scale (PF) of the EORTC QLQ–C30 (version 1.0), which is a linear combination of five binary response items transformed to a 0 to 100 scale, with higher scores representing a higher level of functioning. The data were obtained from the EORTC trial 30893, which was designed as a prospective multicenter randomized phase III study comparing orchidectomy and orchidectomy plus mitomycin C (15 mg/m^2 intravenously every six weeks until progression) in patients with poor prognosis metastatic prostate cancer. One hundred and eigthy nine patients were randomized in the trial. QoL was assessed at six weekly intervals during the first nine months and three-monthly thereafter until progression of disease. A modified version of the EORTC QLQ-C30[1] was used to evaluate QoL.

Most methods based on generalized linear models methodology (i) are useful for both discrete and continuous outcomes, (ii) do not require a constant number of repeated measurements per experimental unit, (iii) allow for differing measurement times across subjects and flexible covariate structures (discrete or continuous, time–independent or –dependent), and (iv) can accomodate missing data (MCAR).

Generalized linear models for longitudinal data can be categorized into three families. Firstly, the generalized linear model can be expressed as a marginal model where the marginal expectation $\mu_{it} = E(y_{it})$ ($i = 1, \cdots, N$ refers to an experimental unit and $t = 1, \cdots, n_i$ refers to a measurement time) is directly modelled in terms of covariates of interest, the marginal expectation being the average response over the subpopulation that shares a common value of the covariate vector. Associations among repeated observations are modelled separately. Secondly, it can be expressed as a random–effects model. Here the outcomes are modelled conditional on an unobserved (latent) random effect or a set of random effects. Subject–specific random effects are assumed to account for all the within–subject correlation that is present in the data. The individual–specific effects are used to explicitely model the heterogeneity among individuals. Thirdly, we mention conditional models in which an outcome is modelled conditional on the other outcomes or at least a set of other outcomes. For instance in transition (Markov) models, the conditional expectation of a current response, given past responses, is assumed to follow a generalized linear model.

In the linear model case, a marginal interpretation can be given to regression coefficients arising from each of the three approaches. However, whenever a nonlinear link function is imposed (e.g., in the case of binary outcome variables), the three approaches give different interpretations for the regression coefficients. More specifically, marginal models are most appropriate for making inferences about population averages. They are often applied in a clinical trial setting, since there the focus is generally on assessing average differences between several treatment arms. Whereas marginal models follow a so-called population–averaged approach, random–effects models adopt a subject–specific approach. In the latter situation, regression coefficients have interpretations in terms of the influence of covariates on both an individual's response and the average response of the population. In transition models, different assumptions about time–dependence, generally imply different interpretations of the regression coefficients.

As outlined in Little's review paper[2], a further classification of models exist into selection models and pattern–mixture models which approach the issue of dropout in two distinct ways: in selection models the dropout probability is conditional on the measurement process, whereas in pattern-mixture models the measurement model is conditional on the dropout pattern. Selection models are widely used but, within the framework of a likelihood analysis, they require relatively detailed specification of the nature of the dropout mechanism. In pattern–mixture models specification of the precise form of the dropout model may not be necessary, unless marginal quantities have to be derived. In this paper we focus on selection models. For additional information on pattern–mixture models, we refer to the papers by Little[2], Hogan *et al.* [3] and Curran *et al.* [4].

The focus of this paper is on marginal models for a binary response. Considering the initial goals of the clinical trial, this type of model is the most reasonable one. In particular, the expectation of the binary response at time t is related to a time trend and a set of covariates by the known linear logistic link function. Various methods for estimating the parameters of these (marginal) models are examined: likelihood based or using alternatives to likelihood theory. Within the likelihood framework, we propose a model which parametrizes the association in terms of marginal odds ratios (Molenberghs and Lesaffre[5]). Alternatively, we estimate the parameters of proposed marginal models by using the generalized estimating equation approach (GEE) and by using weighted generalized estimating equations (WGEE).

References

Aaronson, N.K., Ahmedzai, S., Bergman, B., et al. "The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology," *Journal of the National Cancer Institute, 85, 365 – 376 (1993).*

Little, R.J.A., "Modelling the drop-out mechanism in repeated-measures studies," *Journal of the Americal Statistical Society, 90 nr 431, review paper (1995).*

Hogan, J.W. and Laird, N.M., "Mixture models for the joint distribution of repeated measures and event times," *Statistics in Medicine, 16(1-3), 239-257 (1997).*

Curran, D., Molenberghs, G., Aaronson, N.K., Fossa, S.D. and Sylvester, R.J., "Analyzing longitudinal continuous quality of life data with dropout," Submitted 2000.

Molenberghs, G. and Lesaffre, E.,"Marginal modelling of correlated ordinal data using a multivariate Plackett distribution," *JASA, 89, 633-644 (1994).*

PET Attenuation Correctiom based on Short Transmission Scans

Y. Pawitan, V. Bettinardi, M. Teras
Department of Statistics UCC, Cork
Scientific Institute H. San Raffaele, Milan
TUKS PET Center, Turku, Finland

Abstract

Interactive Clustering: Overview and Tools

Wayne Oldford
University of Waterloo

Abstract

Cluster analysis is typically regarded as a black box algorithm whose output is either a partition of the input data into a number of mutually exlusive groups or clusters, or a sequence of nested partitions which form a hierarchy of clusters. This approach was well suited to the computational resources available when clustering methods were first developed in the 1960s and 1970s but seems inappropriate for today's technology.

Integration with iteractive exploratory data analysis requires the black boxes to be opened and certain key components to be made directly accessible to the analyst.

The approach taken here is that interactive clustering fundamentally involves exploration of the space of possible partitions of the dataset. Movement from one partition to another in this space is carried out by a sequence of moves which can be characterized as either **reduction**, **refinement**, or **reassignment** (the 3 Rs of clustering). Such focus on partitions also opens up the possibility of the development of tools which operate on a collection of partitions.

This presentation gives an overview of this conceptual model and describes an interactive software model which follows it. A prototype of this software model written in **Quail**
(see http://www.stats.uwaterloo.ca/Quail)will be demonstrated to illustrate the kind of interactive clustering tools which are possible.

Mathematical Thinking Versus Statistical Thinking; Redressing the Balance in Statistical Teaching

Michael Stuart
Trinity College Dublin

Abstract

Mathematical thinking tends to emphasise models, methods and procedures and to require development of theory before application in special cases. Problems are formalised as optimisations; solutions are presented as methods to be applied. Modern statistical thinking emphasises processes, notes that all processes are subject to variation, requires data to reflect variation and uses methods for interpreting data in context. Contrasts between the two are drawn out using a series of illustrations and case studies. It is contended that the dominance of mathematical thinking has been largely responsible for the failures of statistical teaching at an introductory level and that an emphasis on statistical thinking in the teaching of our subject will help to improve our success rate.

Promoting Statistical Awareness at Secondary School Level in the Irish Context

Philip J. Boland
National University of Ireland, Dublin

Abstract

The Sixth International Conference on Teaching Statistics (ICOTS-6) will be held during the week of 7 – 12 July 2002 in Durban, South Africa (http://www.beeri.org.il/icots6/). The main theme of the conference is "Developing a Statistically Literate Society". Are we as statisticians in Ireland doing our part to achieve this laudable goal? Can we do more to encourage people in general to use statistical methods and to think statistically?

Statistics forms a very minor part of the Leaving Certificate subject of mathematics in the Republic of Ireland secondary school exam system. Some would say there is in fact no Statistics at all! For example the core of the Higher Mathematics curriculum treats: Fundamental Principles of Counting, Discrete Probability, and Statistics (weighted means and standard deviations). A very small number of students (approximately 5%) also study further probability and statistics as their optional topic (the vast majority take further calculus). This optional topic gives a brief introduction to basic probability, populations and samples, confidence intervals and hypothesis testing. There is little or no emphasis on data analysis (or data handling).

There is a dire need to inform both students and their teachers about the challenging and applicable nature of statistics, as well as how statistical thinking comes into diverse areas of modern life. I believe that a key issue in this context is the local relevance of statistical applications and examples. As part of an overall policy of promoting Science in general in the Republic of Ireland, presentations have been made to secondary school students and teachers highlighting the many uses of statistics and the need for more statistical thinking. In doing so, an emphasis has been put on trying to use examples of local or national interest. My belief is that statisticians in similar situations should endeavour to co-operate in assembling relevant material, which can be used to promote statistical thinking.

With this in mind, I intend to demonstrate this tenet through a series of examples generated for Irish secondary school students. I initially trace the historical origins of probability in games of chance, but quickly move on to highlight how statistical thinking comes into diverse areas of modern life trying to emphasise relevance. The use of statistics in Forensic science is illustrated through an example of probabilistic evidence in the Irish trial for the murder of Lord Louis Mountbatten. Other graphical techniques are also used to highlight the interesting nature of how the number of daily births in Ireland (and I expect most countries) varies by day of the week. The statistical properties of a good diagnostic test are illustrated with reference to the Tuberculin test used to detect Bovine Tuberculosis in Ireland (Bovine TB is a well known and persistent problem in Ireland of considerable economic importance) and plot its geographical variation. These and other examples are meant to highlight how we can try to promote the power and use of statistical thinking.

From the X Files (and Y Files) of the Statistical Laboratory at the University of California, Davis

Francisco J. Samaniego
Professor of Statistics and Director, Statistical Laboratory, UC Davis

Abstract

This presentation will begin with some reflections on how much our subject has changed over the last century. Some of the highlights among the major advances in both theoretical and applied statistics will be emphasized. The question of how these changes have affected the work of the "statistical consultant" will be examined next. Several recent projects "from the files of the Statistical Laboratory at UC Davis" will be discussed as examples. Applied problems in the areas of conservation, traffic engineering and assessment of school effectiveness will be featured. The examples will be used to support the general conclusion: Let's use the modern tools of theoretical, applied and computational statistics to full advantage, **but** let's not throw out the baby with the bath water.

Short-term and long-term variability of standard deviation scores for size in children

Anne Sheehy
Department of Biostatistics, University of Zurich, Switzerland

Abstract

Our primary objective is to quantify long-term and short-term variability in the standard deviation scores (SDS's) for six skeletal size variables and body mass index (BMI) in children and to compare average values of these quantities for boys with those of girls and to make comparisons across variables. The analysis is based on measurements made regularly for 120 boys and 112 girls from 1 month until 20 years for seven variables (standing height, sitting height, leg height, arm length, hip width, shoulder width and BMI) as part of the 1st Zurich longitudinal growth study. Variation in these scores due to variablity in the timing of the pubertal spurt (PS) is seperated out by rescaling the age axis on an individual basis and comparing children with the same developmental age rather than the same chronological age. For a given child, the relationship between the value of its SDS and age is modelled as the sum of an arbitrary (child dependent) smooth function plus an error term. The long-term variability for that child is defined to be the mean square of the departures of this smooth function from its mean level while the short-term variablity is defined to be the variance of the error term. We found that girls' SDS scores have significantly more long-term variability than those of boys, while there is no significant difference between the sexes for short-term variability. Shoulder width, BMI and sitting height have significantly more long-term variation than the other variables. Shoulder width and BMI have the largest short-term variability and standing height has the smallest. Correlations between long-term variablity and adult size and timing and intensity of the PS were small. The results of this analysis are intriguing. Why is the underlying growth process of girls more variable than that of boys? Differences across skeletal parameters are also interesting and deserve further consideration.

The Presentation and Analysis of Bivariate Survival Studies

John Newell
Department of Mathematics, National University of Ireland, Galway

Abstract

Studies involving matched or paired survival data, as particular cases of dependent or bivariate survival data, occur reasonably often in a variety of practical contexts ranging in the authors' experience from ophthalmic surgery to environmental chemistry. This presentation is a follow up to a paper presented at CASI 2000 and attempts to provide additional graphical and inferential techniques to facilitate the analysis of such data based on the bivariate survivor function and various ratios of probability functions.

In addition, a non-parametric approach of estimating the distribution of the difference in survival times within a pair in order to make inferences on an appropriate summary of such a distributional estimate based is discussed.

All the techniques are illustrated on data from a matched study comparing survival in two distinct populations of Malignant Melanoma patients and a paired study comparing time to failure of two Orthodontic brackets.

References

Aitchison, T.C., Newell, J. The Presentation and Analysis of Bivariate Survival Studies. *Statistics in Medicine 2000* (Submitted)

Newell, J., Kay, J.W., Aitchison, T.C. Survival Ratio Plots with Permutation Envelopes in Survival Data Problems. *Statistics in Medicine 2000* (Submitted)

S–PLUS 6 Demonstration

Trevor McMullan
Insightful

Abstract

This brief presentation will highlight some of the new features in S-PLUS 6. In particular two new libraries, Robust Methods and Missing Data methods will be featured. The new GUI (Graphical User Interface) for S-PLUS 6 for Unix and Linux will be demonstrated. S-PLUS 6 for Windows Release Candidate 1 should also be available and some of the new Windows features will be discussed. A brief mention of new server products and a future data mining product will be made. Also reference will be made to a list of new (2000/01) Springer and Wiley S-PLUS publications. Attendees will have the opportunity to try out these new features for themselves during the conference, and to view most of the publications mentioned.

A New Method of Calibration Model transfer in NIR Spectroscopy

J. Huang, D. Brennan, J. Alderman and B. Lane
NMRC, Ireland

Abstract

Introduction

NIR spectroscopy together with multivariate calibration models, as a fast, simple and reliable measurement method, has been widely used on line for process monitoring and control. To construct a multivariate model for prediction of chemical contents, calibration samples are collected with their spectra and corresponding chemical contents being measured. Then calibration methods such as partial least squares (PLS) and ridge regression (RR) can be used to build relationship between spectra and corresponding chemical contents.

Once the model has been developed and set up on line, measurement conditions change may cause the instrument response change such as: replacement of the instrument wholly or part of it, ambient temperature change and raw materiel change in production line. Hence the calibration model may lose its validity. Re-calibrating model is laborious because of the need to base it on a sufficiently broad group of samples.

Recently more efficient methods, known as multivariate standardisation methods, are proposed to transfer the model form one instrument to another. Generally, the procedures are based on the "transfer of spectra ". The methods collect so-called standardisation samples with their spectra being measured on the two instruments. Then compute the relationship between spectra measured on the two instruments. Using this relationship, from spectra measured on the second instrument, spectra on the first instrument can be estimated. Then the model is applied to the estimated spectra to predict chemical contents.

However in many real life problems, as in our EU funded automatic control system (ACS) project, calibration samples (milk samples) are unstable, it is impossible to measure the same samples on two instruments.

Here we describe a novel method of calibration model transfer that does not require measuring the same samples on two instruments.

Methodology

In this paper we focus on calibration model transfer. Hence we suppose a calibration model has already developed for the primary instrument

$y = X\hat{\beta}, \qquad (1)$

where X and y are spectra matrix of calibration samples and corresponding chemical contents vector, respectively, $\hat{\beta}$ is the estimated model coefficients.

To transfer the calibration model to the second instrument (or the second condition), we collect standardisation samples with known chemical contents (y) and spectra being measured on the second instrument (R_2). The transformation matrix F between spectra measured on the two instruments is estimated by minimising the prediction error of applying the model to tranformed spectra

$\|y - R_2 F\hat{\beta}\|$, (2)

where $\|\cdot\|$ is Euclidean distance.

Due to the small number of standardisation samples, minimisation of (2) is always undetermined. Hence some constraints should be imposed on F.

One simple constraint is that the response change can be approximated by one linear function for all wavelengths, i.e., $r_{1,i} = a + br_{2,i}$ and a and b are independent of wavelength i. Then minimisation of (2) becomes minimisation of the following formula

$\|y - (a + bR_2\hat{\beta})\|$. (3)

Let $r_2{}^T$ be spectrum of an unknown sample measured on the second instrument, its chemical property can be estimated by

$\hat{a} + \hat{b}r_2{}^T\hat{\beta}$, (4)

where $\hat{a}$ and $\hat{b}$ are estimated from (3) based on the standardisation samples.

Other constraints on F are also investigated and compared based on data sets generated in our EU funded ACS project.

Conclusion

In this paper we proposed a novel method of calibration model transfer. The method is tested on data sets generated in our ACS project. The method works quite well to transfer calibration models between different measurement temperatures and between different milk types.

Young People, Drug Use and Early School Leaving

Catherine M. Comiskey
Department of Mathematics, National University of Ireland, Maynooth, Co. Kildare.

Abstract

Introduction and Background:

The aim of this talk is to examine the nature and extent of drug use among young people in Dublin and its effect, if any, on the decision to leave school early. In 1995, for the Dublin area, 8.5% (282) of treatment contacts within the Health Research Board were under the age of fifteen when they first used their primary drug of misuse. A further 64.1% (2123) of treatment contacts were aged between fifteen and nineteen when they first used their primary drug of misuse, O'Higgins (1996). Given these data there is a clear and urgent need to examine the extent of hidden drug use among young people and its effect if any on the decision to leave school early.

Methods:

Within this talk our objectives are realised by implementing a statistical methodology known as the capture recapture technique to measure the prevalence of hidden opiate use. The method is applied with data on hospital admissions, police records and methadone treatment from 1996 and with the same data sources (excluding police records) for 1997. The nature of the use of other drugs was examined by conducting a survey amongst 112 early school leavers aged from 14 to 23 years and who had decided to return to education.

Results:

Looking at the 1996 data sources a minimum of 1528 young people aged between 10 and 20 years were identified as using opiates through the 3 data sources. Estimates of the number of hidden opiate users varied depending whether the 2 sample or 3 sample capture recapture method was used. Using three data sources gave a wider and more general definition of opiate use as a non medical source was used. We estimated that approximately 4081 (95% C.I. of 3586 4692) young people aged between 10 and 20 years were using opiates in Dublin in 1996. We found that 51.1% of those surveyed had tried using drugs before they had left school and 73.5% had tried using drugs on or before the age of fifteen. Of those who had tried using drugs before they had left school 46.5% noted that their drug use had effected them at least sometimes in school.

Conclusions:

Finally, while results are interesting and enlightening it must be stated that the methodology is subject to certain assumptions and limitations. The number responding to the survey was modest and one must be careful of drawing general conclusions from a specific study population. In addition we cannot measure the extent to which the assumptions of the capture recapture method are upheld or indeed violated. However, in spite of these limitations results of both known and estimated prevalence of opiate use and the survey results discussed above clearly indicate a need for further study into the links with drug use and early school leaving.

Conditional Phase–Type Distributions and their Application to Geriatric Medicine

Adele Marshall, Sally McClean, Mary Shapcott, Peter Millard
University of Ulster, Jordanstown
Department of Geriatric Medicine, St. George's Hospital, London SW17 0RE,U.K.

Abstract

Previous work has indicated that phase-type distributions are useful in modelling the duration of stay of patients in departments of geriatric medicine. The inclusion of patient details and admission information is also of benefit to the modelling process, in particular the incorporation of any causal information that may exist in the data set. Bayesian belief networks (BBNs) are statistical graphical models ideal at representing causal information by providing a framework for describing and evaluating inter-relational variables. The work in this paper reports on the development of a conditional phase-type distribution (C–Ph), a special kind of phase-type distribution that uses BBNs to represent the causal relationships between patient variables and phase-type distributions to model the continuous distribution of patient survival in hospital.
Keywords. Conditional phase-type distributions, Bayesian belief networks, patient survival.

Introduction

Phase-type distributions describe duration until an event occurs in terms of a process consisting of a sequence of latent phases the states of a latent Markov model. For example, duration of time in hospital can be thought of as a series of transitions through phases such as: acute illness, intervention, recovery, discharge (Faddy, 1994). Bayesian belief networks (BBNs) are special cases of probabilistic graphical models that use directed arcs exclusively to form a directed acyclic graph (DAG) and Bayes Theorem to represent the probabilistic relationships between variables (Buntine, 1996). One of the key features of BBNs is the ability to describe causal relations. Causality is often interpreted as a type of dependency or the effect or direct influence of one variable on another.

The extension of Bayesian belief networks to include continuous variables has been considered by Lauritzen et al. (1989) who introduced Conditional Gaussian (CG) distributions. However the CG distributions are not appropriate for modelling the survival of patients in hospital due to the skewed nature of the data set. Therefore the aim of this paper is to introduce a more appropriate distribution called conditional phase–type distributions. The conditional phase–type (C–Ph) distributions can represent a continuous non–normal variable while also incorporating the causal information in the form of a BBN (Marshall et al., 2000).

The application considered is the modelling of patient duration of stay in hospital using the 'CLINICS' data set, containing 4722 patient records collected between 1994-1997.

Model

The conditional phase-type distribution is represented by a BBN of interrelated causal nodes which precede the effect node or process model. The effect node here is characterised by a continuous positive

random variable, the duration, described by a phase-type distribution. The Causal Network is modelled as a Bayesian belief network and the Process Model defined as a phase–type distribution. The C–Ph model may then be defined as comprising of causal nodes $\mathbf{C} = \{C_1, \cdots C_m\}$ and process nodes $\mathbf{Ph} = \{Ph_1, \cdots Ph_{n.}\}$. A special case of the C–Ph model is the inclusion of an outcome node **O** which acts as a connecting variable between the causal and process models. The parameters of the process (the phase–type distribution) are conditional on the outcome which is the final node in the BBN. The joint distribution of **C** may be represented by:

$$P(\mathbf{C}) = \prod_i P(C_i \mid pa(C_i)) \qquad (1)$$

where pa is the parent set of C_i and the distribution of the process nodes **Ph** is represented by the p.d.f.:

$$f(t \mid \text{pa(process)}) = \mathbf{p} \exp\{\mathbf{Qt}\}\mathbf{q} \qquad (2)$$

where pa(process) are the causal nodes which are parents of the process and the values of **p, Q** and **q** are : $\mathbf{p} = (1\ 0\ 0 \cdots 0\ 0), \mathbf{q} = -\mathbf{Q1} = (\mu_1 \mu_2 \cdots \mu_n)^{\mathrm{T}}$ and

$$\mathrm{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 \cdots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 \cdots & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 \cdots & 0 & -\mu_n \end{pmatrix},$$

Application to Geriatric Medicine

The CoCo and BIFROST packages (Badsberg, 1992) may be used together to determine a suitable graphical structure to represent the associations between variables in a BBN. When the topology of the causal network is known a priori the C–Ph distribution may be obtained by using maximum likelihood methods to fit parameters to phase–type distributions for each set of values in pa(process). MATLAB software (MATLAB, 1992) is used to implement the Nelder–Mead algorithm (Bunday, 1984) to perform the likelihood ratio tests which in turn determine the optimal number of phases in the distribution (Faddy and McClean, 2000). The causal network, produced by BIFROST (p-value=0.01), is incorporated into the phase-type distribution to represent the C-Ph model for the geriatric data (Figure 1).

Figure 1. C–Ph model applied to geriatric medicine

The outcome node of the resulting BBn can relate directly to the process model, the phase–type distribution for duration of stay, to form the conditional phase–type distribution for the Clinics data set.

Summary

This paper has 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. The resulting C–Ph model uses discrete variables for the causal model and a continuous variable for the stochastic process. A special case of the C–Ph model is considered where the outcome node belonging to the BBN can relate directly into the phase–type distribution for the length of stay of patients in hospitals. This simplifies the calculations for patient duration of stay as it is the only node that relates directly in the phase–type distribution therefore allowing the process model, the duration of stay to be locally independent of the causal network. It is therefore easier to calculate the possible duration of stay of an elderly patient into hospital given some patient information.

References

J.H. Badsberg (1992). A Guide to CoCo - An Environment for Graphical Models. Aalborg University, Denmark

B.D. Bunday (1984). Basic Optimisation Methods. Edward Arnold Publishers Limited - London.

W. Buntine (1996). A Guide to the Literature on Learning Probabilistic Networks from Data. *IEEE Transactions on Knowledge and Data Engineering, 8(2) 195-210.*

M. Faddy (1994). Examples of fitting structured phase-type distributions, *ASMDA, 10, 247-255.*

M. Faddy, S. McClean (2000). Analysing Data on Lengths of Stay of Hospital Patients Using Phase-Type Distributions. *Applied Stochastic Models and Data Analysis.*

S.L. Lauritzen, N. Wermuth (1989). Graphical models for Associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics 17, 31-57.*

A.H. Marshall, S.I. McClean, C.M. Shapcott and P.H. Millard (2000). Learning Dynamic Bayesian Belief Networks using Conditional Phase-Type Distributions. *LNAI 1910, PKDD 516-523.*

MATLAB (1992). Reference Guide, The MathsWorks Inc., Natick, Massachusetts.

Spurious and Non-Spurious Power– the Grubbs' Outlier Test Case

Kevin Hayes, Tony Kinsella
University of Limerick

Abstract

The treatment of an outlier test in an introductory statistics course provides an opportunity to motivate a student's thinking on type II error and power where both of these measures contain spurious and non-spurious components. The division of the type II error and power into these components arises in the context of outlier tests, as it is not necessary that a contaminant observation be identified as the outlier when applying such tests. This suggests that the power function is not the most appropriate test performance criterion to use. Alternatives are considered, including a new performance criterion not encountered in the literature. The learning benefit for the student is a requirement to think carefully about criteria suitable for measuring the performance of statistical tests. The aim of this paper is to direct students' attention from a pre–occupation with searching for statistical significance in data and re–emphasize these more important aspects of hypothesis testing.

Curve Registration and Alignment

William Ryan
University of Limerick

Abstract

TEXT

Text

Mixture Models for Ranking Data

Brendan Murphy, Donal Martin
Trinity College Dublin

Abstract

Ranking data arises when judges are asked to rank some or all of a group of objects. Examples of ranking data arise in the Irish electoral system where voters rank some (or all) of the candidates, and the Irish College admission system where prospective students rank up to ten courses from a list of approximately three hundred courses.

Each ranking consists of a permutation of some (or all) of the objects under consideration. The standard methods of data analysis are not appropriate for this data.

Various statistical models for ranking data have been prosposed in the literature. We propose using mixture models to model heterogeneous populations of judges. Some methods for fitting such mixtures will be demonstrated.

Analysis of the CAO Database

Clare Crinion
Trinity College Dublin

Abstract

The Central Applications Office (CAO) processes applications to third level colleges in Ireland. Set up in 1976, it now processes more than 50,000 applications each year. The analysis carried out in this study is based on applications to degree courses only and includes data from the years 1996 to 2000. A number of data mining techniques have been used to identify patterns in the data, mainly clustering and association rules. Multi-dimensional scaling has been used as a visualisation technique to plot the clusters.

An interesting feature of the data is the fact that students choose courses based on preferences. Courses chosen are rated 1 to 10. The research has tried in some part to address this feature. The outcome gives quite interesting insight into how students choose their courses and some of the factors, which influence their decision.

Ordering variables in Displays of Multivarate Data

Catherine Hurley
National University of Ireland, Maynooth

Abstract

Scatterplot matrices and parallel coordinates plots are two methods of visualising multivariate data. Both become less effective as the number of cases and variables increase, presenting us with an overwhelming amount of information that can be difficult to absorb. In this presentation we examine ways of ordering variables so that multivariate data displays are more easily comprehended.

According to Hills(1969) the first and sometimes only impression gained from looking at a large correlation matrix is its largeness! The same accusation could be levelled at scatterplot matrices. Bertin(1983) advocates the use of diagonalisation to simplify diagrams. For scatterplot matrices this suggests that we permute the variables so that the biggest (absolute) correlations are close to the diagonal with values decreasing as one moves further from the diagonal. There is a vast literature on this so-called object seriation problem, see for example Kendall (1971) and Hubert (1974). While finding the best permutation requires a permutation search, more efficient algorithms based on minimal spanning trees, hierarchical clustering or multi-dimensional scaling provide orderings which are sufficient for our purposes.

The scatterplot matrix with the permuted variables should be easier to interpret because panels of similar variables appear together in a block. We also enhance the scatterplot matrix by using different background colours in the panels for various levels of correlation.

Parallel coordinate plots are relatively immune to the curse of dimensionality, by comparison with scatterplot matrices. However, with moderate numbers of cases the plots become cluttered and it is hard to see any pattern except for very obvious clusters and outliers. Generally, patterns are easier to discern when successive variables are (positively) correlated because this minimises the number of crossings. This suggests that we permute the variables so that each variable is adjacent to the variables with which it has the highest correlations. Similarly, we could search for the best path through the variables using travelling salesman methods. While the optimal path is not generally found in polynomial time, various heuristic methods such as those suggested for scatterplot matrices provide orderings which yield improved parallel coordinate displays.

References

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps*, translated by W.J. Berg, University of Wisconsin Press.

Hills, M. (1969). On looking at large correlation matrices. *Biometrika*, 56,249-253.

Hubert, L. (1974). Some applications of graph theory and related non-metric techniques to problems of approximate seriation: the case of symmetric proximity measures. *British Journal of Mathematical and Statistical Psychology* 27, 134-153.

Kendall, D.G., (1971). Seriation from abundance matrices. In F.R. Hodson et al (eds) *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press.

The Use of Weighted Census Based Deprivation Indices in Small Areas

J.A. Saunders
Centre for Health Services Studies, University of Kent, Canterbury, Kent

Abstract

Backround

The concept of deprivation refers to the conditions experienced by people who are poor while the concept of poverty relates to the lack of income and other resources that make these conditions difficult to break away from. The usual census based deprivation indices are often based on equally weighted (and highly correlated) variables. This is in spite of the fact that different social groups have been shown in numerous studies to have differing probabilities of suffering from deprivation. A weighted deprivation index based on individual level data and Census data produces a more accurate and more easily understandable method of estimating deprivation within an area as it reflects the differences between social groups. This paper will describe how a methodology used by Gordon (1995), involving the calculation of weighted census based deprivation indices, can be applied to smaller areas.

Method

Survey data from a South London Borough and a South East Health Authority area were used together with Census data (OPCS) to estimate the percentage of poor households in wards using similar methods to those used when estimating the number of deprived nationally using the 'Breadline Britain' survey findings. The surveys included questions on health and lifestyle of the sampled population together with questions on problems faced in everyday life and whether they lacked any 'essential' items from a list derived from earlier studies, Gordon and Pantazis (1997). The deprivation indicator of lack of 3 or more essential items was used as the dependent variable in a logistic regression analysis together with variables often used in deprivation indices to try to predict their relative weightings for use in estimating the number of poor households in an area. Since the variables were chosen so they could be used in an additive manner the number of households calculated for each variable could be totalled to form an estimate of the total number of poor households.

Results

The results produced in both areas were very similar to the numbers estimated using the Breadline Britain weightings but (not surprisingly) there were differences in individual weightings for some of the variables.

Conclusion

This methodology can be applied to smaller areas to give estimates of the number of deprived using the nationally derived weightings. More accurate local estimates, subject to the different local conditions, can be easily derived if a similar survey to the Breadline Britain survey is conducted locally.

References

Gordon, D. (1995) Census based deprivation indices: their weighting and validation. *Journal of epidemiology and Community Health, 49(Suppl 2): S39-S44.*

Gordon, D. and Pantazis, C. (eds.) (1 997) *Breadline Britain in the 1990's*, Ashgate Publishing Ltd: Aldershot.

Modeling Hepatitis C Dynamics in Dublin's Intra–Venous Drug Users

Declan Walsh
Department of Mathematics, National University of Ireland, Maynooth, Co. Kildare

Abstract

With the advent of blood screening the primary mode of transmission of HCV has been via Intravenous Drug Use (IDU). It has been estimated that the proportion of HCV infection linked to IDU may be as high as 75%. It has also been shown that syringe exchange programs have had little impact on the prevalence of HCV among IDUs with estimates of HCV prevalence in Dublin's IDU population ranging from $51.2\% - 76\%$.

Treatment of chronic HCV has also proven to be expensive. It has been estimated that the discounted cost per lives save for treatment with interferon–alpha ranged from $£2,142 - £17,128$. Similar results have been found by Shiell A. et al.(1999) with discounted costs per lives save estimated at $19,110. Also the effectiveness of interferon alpha treatment is poor with 12 months of treatment leading to a sustained virological response in the range 20-40%, though combined treatment with Ribavirin yields better rates.

Furthermore the estimated long terms economic burned on governments is set to increase with estimates by Wong et al.(2000) indicating that HCV in the U.S. is likely to cost that state 10.7 billion in direct medical expenditure over the period 2010–2019 and a cost in societal terms estimated at 86 billion dollars.

With these consideration in mind it is proposed that a model of the dynamics of HCV in Dublin's IDU population be constructed. From this model it is hope that such issues as long term cost, optimal reduction strategies and dynamics behavior of the disease be determined.

Deterministic HCV models to date have not considered specifically how transmission within an exclusively IDU population may come about. It is hoped that this problem may be surmounted by incorporating user-needle contact assumptions already applied in the modeling of HIV dynamics within IDU populations along with natural history consideration exclusive to HCV, such as a four-stage disease progress and gender and age dependent progression.

References

Shiell, A., Brown, S., Farrell, G.C. Hepatitis C: an economic evaluation of extended treatment with interferon. *Med J Aust.1999 Aug 16; 171(4) 189-93.*

Wong, J.B.; McQuillan, G.M.; McHutchison, J.G.; Poynard, T. Estimating future hepatitis C morbidity, mortality and cost in the United states. *Am J Public Health Oct; 1999 90 (10): 1562-9.*

Analysis of Multivariate Measurements of Rowing Biomechanics

F. O'Sullivan, J. O'Sullivan, M. Bradley, M. Kenneally
Department of Statistics, University College Cork

Abstract

Rowing is a sport that is growing in popularity and continues to make strides in technique improvements. With the increased competitiveness comes an increased rate of injuries, most of which are to the lower back. A study was recently published (Measuring spinal motion in rowers: the use of an electromagnetic device, Anthony M.J. Bull and Alison H. McGregor, Clinical Biomechanics 15, 2000, 772-776) in which the authors made dynamic measurements of spinal and pelvic movement during use of an ergometer (rowing machine) during a rowing session. The subjects were rowers from the Imperial College boat club; they were asked to simulate a number of 'poor' techniques along with their usual style of rowing. The researchers used an electromagnetic device (Flock of Birds) to assess spinal kinematics during the rowing sessions to measure multiple areas of the spine continuously. Their objective was to determine whether the measurement technique would allow discrimination between rowing techniques that would affect performance and injury rate.

We have acquired a segment of this data, and have been working to develop the statistical techniques that could be used to analyse this, and further data. The data we have used consists of measurements on one individual over 4 two-minute sessions. The variables studied were 'force', and 15 different components of back displacement and motion. Our initial approach has involved using principal components to i) reduce the dimensionality of the data, and ii) generate a model for the stoke level data, so that parameters can be studied in relation to variations in technique and ultimately used to correct faults in style.

The figures below are generated using the force variable. Figure 1 shows the force profile for the 139 strokes available for the rower. The first 30% of the stroke is where all force or pull is exerted, and the remaining 70% is 'recovery' (returning to the forward postion). Figure 2 is the scree plot for the principal components analysis of the 'force' data. The first three eigenvectors are shown in figure 3, and figures 4–6 show the coefficients for the first three principal components. In the coefficient plots one can see that the four different sessions (illustrated by colour) appear to have different characteristics. In the presentation we present corresponding analyses for the displacement data and show how the results

provide a simple empirical model for describing key aspects of rowing biomechanics on the ergometer.

We wish to thank Drs Alison McGregor, Anthony Bull and those in their laboratory for allowing us the use of these data.

Choosing Optimum Subset of Wavelength in Near Infrared Spectroscopy

Arief Gusnanto
University College Cork

Abstract

Near Infrared (NIR) Spectroscopy is a method for estimating the concentration of chemical content such as fat or protein based on reflectance of infrared radiation. It is cheaper but less accurate compared to standard chemical analysis. Before it is used, the instrument of NIR needs to be calibrated. In calibration of NIR instrument, we estimate the relationship between a known concentration of a chemical compound as a response variable, and its reflectance spectra as the explanatory variables. The reflectance spectra from each observation are measured over a range of wavelengths. The number of wavelengths available for analysis is often very large and it is not clear whether all is needed for prediction. The objective of this study is to choose optimum number of wavelengths to be put into the prediction model. A team in National Microelectronics Research Centre (NMRC), Ireland, under Automate Control System (ACS) project, collected the data for the analysis. There are 70 milk samples drawn from a milk factory, with spectra from 130 wavelengths ranging from 829 to 1145 nm. The response variable is concentration of fat in the milk. For the purpose of calibration, we split the observations into two subsets. The estimation set, randomly selected from the samples, is used to estimate the parameter of ridge regression (RR) and partial least squares (PLS) regression, and obtain optimum configuration of wavelengths based on variable selection. The remaining observations, the validation set, are used to measure prediction accuracy of the model and to choose the optimum subset of wavelengths.

For ridge regression, the result shows that less than 20 wavelengths are sufficient to make the prediction. In conclusion, not all of the available wavelengths are necessary for optimum prediction in NIR spectroscopy; in fact, only a fraction of them is needed. Achieving a small number of wavelengths is of practical interest since it leads to a simpler and faster computation.

KEYWORDS: Calibration, variable selection, ridge regression, partial least squares regression

Application of Nonparametric Regression Methods to Ridge Parameter Estimation

M. Byrtek, F. O'Sullivan
Department of Statistics, University College Cork

Abstract

Disease Mapping in Ireland: Current Databases and Mapping

Kathleen O'Sullivan
Statistical Laboratory, University College Cork

Abstract

Maps have been used to illustrate the geographical distribution of disease at local, national and international scales for over a century. Howe (1989) provides a detailed discussion on the evolution of such maps. The earliest examples of disease mapping originate at the end of the 18th century when spot maps were used to illustrate infectious diseases such as yellow fever. The earliest surviving multicoloured map, which depicts the distribution of an infectious disease, comes from the City of Glasgow and this map highlighted the finding that the condition was most prevalent where housing was poor and crowded (Kemp et al., 1985). In the past, cancer maps produced showed area based mortality rates where mortality was often taken to provide a satisfactory approximation of incidence. Haviland (1875) produced the first map of cancer that displayed degrees of cancer mortality using tints of blue and red. In the mid 1980s, maps were produced based on incidence data (Kemp et al., 1985; Carstensen and Jensen, 1986). Kemp et al. (1985) argued the value of using incidence over mortality data in representing temporal and spatial patterns of cancer. They produced the first cancer atlas based on incidence data using the 56 administrative units of Scotland.

Mapping of disease rates enables the spatial patterns across a country to be visually displayed. These spatial patterns could include clusters of neighbouring areas with similar rates or gradients of rates. By using maps, areas of low and high rates can be compared with respect to their environment and other characteristics. This may lead to the development of hypotheses about causal factors, which may be responsible for any differences (Gardner et al., 1983). The comparison of areas with high and low rates and the identification of areas of high risk of cancer may identify factors, which influence the risk of cancer. The mapping of the variation in cancer rates provides a mechanism for evaluating observations of local clusters of cancer cases (Carstensen and Jensen, 1986).

The analysis of the distribution of disease has often led to its cause and ultimate prevention. A classical example is that of Snow (1854), where he recorded and mapped the addresses of victims of a cholera epidemic in London. He showed that the epidemic of cholera in London could be related to the pollution of the public water supply. English (2000) and Carstensen and Jensen (1986) provide further examples.

The variation in disease rates has been studied in detail using large geographical areas (Cuzick and Elliott, 2000). Examples include regions in England and Wales, municipalities in Finland, and counties in the USA. In contrast little is known about the variation in rates over small areas (Cuzick and Elliott, 2000). Cuzick and Elliott highlight the importance of studies at small areas level as studies using higher levels of aggregation can miss important local disease clusters. For example, childhood leukaemia in Britain shows regional and sub regional variation in incidence rates. At district level known local areas of high incidence, for example Sellafield, fall well within the observed distribution of incidence rates for Britain as a whole (Stiller et al., 1991).

In the Irish context, standardised age-adjusted incidence rates for various cancers have been illustrated using maps for small geographical areas within the Cork and Kerry region (Crowley, 1995). The Irish National Cancer Register produced maps of the standardised incidence ratio (SIR) for different cancers at county level (National Cancer Register, 1997; 1998). Maps of standardised low birth weight incidence ratios have been produced for County Dublin and standardised mortality ratios for asthma nationally (Kelly, 1999).

This paper will present a review of disease mapping focusing on cancer mapping using illustrations from published material. It will discuss potential databases such as the National Cancer Registry. It will illustrate the use of a mapping tool, MapInfo, to produce maps based on data from the National Cancer Registry.

Keywords Maps, disease mapping, cancer mapping, MapInfo

References

Carstensen B. and Jensen O. (1986). *Atlas of Cancer Incidence in Denmark 1970-79. Denmark: Danish Cancer Registry.*

Cuzick J. and Elliott P. (2000). Small-area studies: purpose and methods. *In: Elliott, P., Cuzick J., English D. and Stern, R. (Eds), Geographical and Environmental Epidemiology.* Oxford: Oxford University Press.

Crowley M.J. (1995). *Cancer - The Irish Experience. Cork: Statistical Laboratory, University College Cork.*

English D. (1999). Geographical epidemilogy and ecological studies. *In: Elliott, P., Cuzick J., English D. and Stern, R. (Eds), Geographical and Environmental Epidemiology. Oxford: Oxford University Press.*

Gardner M.J., Winter P.D., Taylor C.P., Acheson E.D. (1983). *Atlas of Cancer Mortality in England and Wales 1968-1978. New York: John Wiley and Sons.*

Haviland A. (1875). *The geographical distribution of heart disease and dropsy, cancer in females and phthisis in females in England and Wales. London: Swan Schonnenschein.*

Howe G.M. (1989). Historical Evolution of Disease Mapping in General and Specifically of Cancer Mapping. In: Boyle P., Muir, C.S. and Grundmann E. (Eds), Cancer Mapping. *Recent Results in Cancer Research, Vo 114. Berlin: Springer-Verlag.*

Kemp I., Boyle P., Smans M., and Muir C. (Eds). (1985). Atlas of Scotland 1975-1980: Incidence and epidemiological persepective. *IARC Scientific Publications No. 72. Lyon: International Agency for Cancer Research.*

Kelly A. (1999). Case Studies in Bayesian Disease Mapping for Health and Health Service Research in Ireland. In: Lawson A., Biggeri A., Bohning D., Lesaffre E., Viel J.F. and Bertollini R. (Eds), *Disease Mapping and Risk Assessment for Public Health.* New York: John Wiley and Sons.

National Cancer Registry. (1997). *Cancer in Ireland, 1994.* Cork: National Cancer Registry.

National Cancer Registry. (1998). *Cancer in Ireland, 1995.* Cork: National Cancer Registry.

Snow J. (1854). *On the mode of communication of cholera (2nd edn).* London: Churchill Livingstone.

Stiller G.A., Draper G.J., Vincent T.J. and O'Connor C.M. (1991). Incidence rates nationally and in administratively defined areas. In: Draper G. (Eds), *Geographical epidemiology of childhood leukaemia*

and non-Hodgkin's lymphoma in Great Britain, 1966-83. Studies on Medical and Population Subjects No 53. London: HMSO.

Size and Shape Analysis of the Human Mandible–age 9 to 15 Years

Valerie Easton, John McColl
ICON
Department of Statistics, University of Glasgow

Abstract

Introduction

In the field of dentistry, there is much to be gained by numerical description of complex forms, like the cranio–facial complex. The way in which the mandible grows i.e. changes in size and shape over time, is of great importance in many branches of dentistry, particularly orthodontic dentistry. Mandibular bone growth is a complex process and it is the job of the orthodontist to ensure that this growth is as 'normal' as possible. It is the job of an orthodontist to monitor and predict any problems that might occur during normal growth and development. If a baseline of data on 'normal' growth changes of the mandible during a child's development could be identified, the orthodontist would be able to recognise abnormalities during the early stages of a child's development and tailor any treatment plan that may be required for individual patients. The mechanics of size and shape analysis would allow numerical representation of a baseline of 'normal' bones and any improvement attributed to treatment could be compared to such a baseline of data on 'normal' growth changes. An orthodontist (or indeed anthropologist) might also be interested in the change over time of the size and shape of the bone and such observations would require quantification of the size and shape of a series of mandibles. They might also wish to examine whether there are any size and shape differences between males and females, or differences between different ethnic groups. Again such comparisons would require quantification of the size and shape of the mandible for accurate description, or indeed series of mandibles to monitor size and shape changes over time between different groups. In the past, there have been many attempts to capture the size and shape information inherent in complex irregular objects by numerical representation. One such method that has been used with much success is that of Procrustes Analysis (Dryden IL and Mardia KV, Statistical Shape Analysis, Wiley, Chichester 1998).

Data

The data sample available for investigation consists of a longitudinal series of lateral head cephalograms from the BC Leighton Growth Study, which was started in the early 1950's at Kings College School of Medicine and Dentistry, London. A sample of 84 subjects who had each been x-rayed annually from around the age of 2 years, through to the age of 20 years were made available. A subset of 23 subjects, with 3 available films for ages 9, 11, 13 and 15 was selected following certain inclusion and exclusion criteria. The mandibular data was prepared for subsequent use with the method of Procrustes by carefully tracing and identifying anatomical and intermediate landmarks around each outline then digitising each of these landmarks to produce a computerised set of (x,y) co-ordinate points which characterise each mandibular outline in the sample.

Method

Procrustes analysis can be thought of as a homologous point technique that provides us with the tools to describe size and shape information of any complex form in mathematical terminology, as well as allowing comparison between forms. Consider two forms or configurations of homologous landmark points of two complex outlines i.e. a pre–assigned correspondence between the points of the two configurations. Procrustes analysis allows comparison of two such forms by matching the two forms with similarity transformations of translation, rotation / reflection and scale change to be as close as possible according to Euclidean distance using least squares techniques. Since 'shape' is defined to be that which remains after differences in location, orientation and scale have been removed from forms we have such an optimal matching method in Procrustes analysis. Procrustes analysis is applied to the sample of mandibular outlines in order to investigate its usefulness in describing the size and shape of the human mandible for ages 9, 11, 13 and 15 years, for both males and females. Differences between the size and shape of the mandibular outlines for males and females are also assessed, for each age Shape variability within samples is also explored by way of principal components analysis.

Results

Overall, the mandible was observed to be 'growing' between ages 9 and 15 i.e. changing in both size and shape over a period of time. There was no difference in terms of the size and shape of the bone between males and females in the sample, for each age. In addition, investigating shape variability by way of principal components analysis, resulted in broadly similar patterns for males and females at different age groups of 9, 11, 13 and 15 years.

Conclusion

Procrustes Analysis provides a very useful and practical framework in which to describe the size and shape of complex irregular forms like the mandible.

A New System of Consumer Demand Equations

Denis Conniffe
The Economic and Social Research Institute, Dublin

Abstract

The choice of a set of equations to describe any system–physical, biological or economic–must depend on what we know, or theorise, about the system as well as on purely statistical criteria of goodness of fit to data. There are few topics in economics where the theoretical requirements are more constraining than in specifying consumer demand equations. These relate quantities of commodities purchased to prices and income. The economic theory of a consumer's behaviour suggests, or perhaps imposes, substantial constraints - aggregation, homogeneity, Slutsky symmetry and negativity - on the equations. The terms may be unfamiliar to non-economists, but at least some constraints are easily understood. The aggregation restriction means that expenditures on goods must add up to the consumer's budget, or "income", (borrowing/saving can be treated as a commodity priced by the interest rate); the homogeneity restriction means that if income and all prices double, nothing changes. Finding models that satisfy the constraints as well as having other desirable properties, such as the ability to represent a wide range of consumer behaviour while being reasonably parsimonious in unknown parameters, is not easy.

Despite much theoretical research and the huge literature it has generated, the choice among valid models is still quite limited. This would not matter much if the available models were truly of wide applicability and acceptability, but most are not. For example, the famous and frequently employed Stone-Geary linear expenditure system (e.g. Neary, 1997) conforms very well with theory and is parsimonious with parameters requiring estimation, but is really only appropriate for relatively few and broadly defined commodities. But we often want a system capable of representing consumer behaviour for a fine breakdown of commodities as regards substitutability and complementarity of various goods. However, models often advocated as particularly appropriate for a large set of specific commodities, for example, the original (now called Fractional) Translog and the Generalised Leontief (e.g. Deaton, 1986) are very demanding on data. This is both because a lot of parameters are involved and because some theoretical constraints (negativity conditions) are not actually guaranteed unless estimated coefficients stay within certain ranges. Indeed, some purportedly powerful systems such as the Almost Ideal Demand System, or the current Translog, are quite fundamentally flawed both in terms of economic theory and statistical specification.

So there is scope for any new system of demand equations that complies with theory, is flexible enough to represent most consumer behaviour and is sparing in unknown coefficients. This conference paper presents such a system. Compliance with theory is assured by commencing from the explicit (indirect) utility function

$$U = \frac{y}{p}\left\{1 - \sum \gamma_j \left(\frac{p_j}{y}\right)^{\beta_j}\right\},$$

where y is income, P is weighted geometric mean of prices, so that $logP = \sum(\alpha_j logp_j)$ with the positive α_j adding to unity, and summations are over the n commodities. U is a valid indirect utility function if it homogeneous of degree zero in income and prices ($\mathbf{p}$), nondecreasing in y, nonincreasing in $\mathbf{p}$, and convex,

or quasi-convex, in **p** and these conditions hold if each $\gamma_i\beta_i$ is positive and $\beta_i < 1$. Then deriving the demand equations from the utility function by Roy's identity guarantees a system satisfying the required constraints. In budget share form, these are

$$W_i = \frac{\alpha_i \left\{1 - \sum \gamma_j \left(\frac{p_j}{y}\right)^{\beta_j}\right\} + \gamma_i\beta_i \left(\frac{p_i}{y}\right)^{\beta_i}}{1 - \sum \gamma_j(1-\beta_j)\left(\frac{p_j}{y}\right)^{\beta_j}},$$

where $W_i = p_i q_i / y$, with q_i being the quantity demanded. Taking all the $\beta_i = 1$ gives the linear expenditure system. The LES is sometimes interpreted in textbooks (e.g. Deaton & Muellbauer, 1980, p. 145) as giving anyone's demands as weighted averages of 'rich' and 'poor' persons' demands, the weights being discretionary and subsistence incomes. Here the demands can also be seen as weighted averages, now from a 'rich' person and from someone obeying Houthakker's (1960) more general indirect addilog system. Now each of these systems is of limited applicability individually. Rich person (Bergson) demands imply unit price and income elasticities, while the indirect addilog system implies that the cross-price elasticities for commodity j with respect to price k are the same for all j. But this weighted average is remarkably versatile as can be demonstrated by examining income, own-price and cross-price elasticities and showing their flexibility. In particular, the Engel curves of the system (the relationships of consumptions to income at fixed prices) can take quite a range of non-linear shapes, which are required to match the empirical findings of many studies (e.g. Lau, 1986).

The system is quite parsimonious in involving only $3n - 1$ parameters as compared to $O(n^2)$ for the other models in the literature that claim flexibility. There are other systems as parsimonious, including a generalisation of the indirect addilog (Barten, 1977), but they cannot model as broad a range of consumer behaviour.

There is some price to be paid for the model's virtues in that estimation is not computationally simple. It is clear the demand equations are very non-linear in the parameters and the systems estimation required having added stochastic components to the model is complicated. However, it is not more complex in this respect than some other models, for example, the fractional translog system. In any case, the computing power associated with modern personal computers has made issues of computational difficulty far less important than they once were.

References

Barten, A.P. 1977. "The Systems of Consumer Demand Functions Approach: A Review", *Econometrica, 45, 23-51.*

Deaton, A. 1986. "Demand Analysis", in Griliches, Z. and M. D. Intriligator (eds.), *Handbook of Econometrics Vol. 3, Amsterdam: North-Holland, pp. 1767-1839.*

Deaton, A., Muellbauer, J. 1980. *Economics and Consumer Behaviour, Cambridge: CUP.*

Houthakker, H.S. 1960. "Additive Preferences", *Econometrica, Vol. 28, pp. 244-256.*

Lau, L. J. 1986. "Functional forms in econometric model building", in Griliches, Z. and M. D. *Intriligator (eds.), Handbook of Econometrics Vol. 3, Amsterdam: North-Holland, pp. 1515-1566.*

Neary, J. P. 1997. "R. C. Geary's Contributions to Economic Theory", in Conniffe, D. (ed.), *Roy Geary, 1896-1983: Irish Statistician, Centenary Lecture by John E. Spencer and Associated Papers, Dublin: Oak Tree Press, pp. 93-118.*

Foreign Direct Investment in the European Union and the Coporate Tax rate

Margaret Hurley
Department of Economics, NUI, Maynooth

Abstract

Ireland is very successful at attracting inward foreign direct investment, particularly from the United States. This has been part of the Irish growth story of the last decade. Ireland's corporate tax rate, at 10% for manufacturing industry, is very attractive compared with that of the other EU countries. This has lead to worries within the EU of 'harmful tax competition'; countries within the EU competing to attract capital and a business activities by lowering taxes, against the spirit of a single European market. However, Ireland has advantages other than its tax rate; it is English speaking, strongly pro-Europe, and up to recently had an excess supply of labour, as measured both by the unemployment rate and the growth of the labour supply. This paper explores the issue of the importance of the corporate tax rate in determining FDI flows, when all of these other factors are taken into account.

A gravity model approach is taken to answering this question. Gravity models are predominantly empirical models that explain connections between markets and countries using geographical variables such as distance and size, with dummy variables included for common language, membership of the same trading bloc and historical links. The most frequent application of these models is in explaining trade flows; other applications included equity flows, migration flows, business cycle and stock market correlations.

$$\ln(FDI)_{ij} = \beta_0 + \beta_1 \ln(GCD)_{ij} + \beta_2 \ln(Size)_j + \beta_3 \ln(Size)_j + \beta_4 Border_{ij} +$$

$$\beta_5 EMU_j + \beta_6 Language_{ij} + \beta_7 Taxrate_j + \beta_8 labourcost_j + \varepsilon_{ij}$$

The model is estimated by ordinary least squares; in the equation above, 'j' refers to the EU country receiving the flow, and 'i' is the source country for the investment. Since data for Luxembourg is included in that of Belgium, there are fourteen 'j' countries. The number of source countries varies for each EU country, as zero or small flows are not included. Data is taken from a Eurostat publication and the most recent year available is 1998. GCD refers to the great circle distance between financial centres of the source and EU country and size is measured by GDP. The border and language dummies are one if the countries have a border or language in common and zero otherwise. The EMU variable is a one/zero dummy capturing whether or not the country was preparing to fix its exchange rate; it is zero for the UK, Denmark, Sweden and Greece and one for all the other countries. It is hypothesised that, all other things being equal, EMU membership should lead to an increase in FDI. The labour cost variable gives the hourly labour costs in euros for each of the recipient countries; this would be expected to have a negative influence on FDI. The variable of most interest is the tax; this is the corporate tax rate for manfacturing industry in the EU–14 countries.

A Co-integration Analysis of Balance of Payments Data

Patrick Murphy
Department of Statistics, University College Dublin

Abstract

The Balance of Payments (BOP) is a statistical statement that systematically summarises, for a specific time period, the economic transactions of an economy with the rest of the world. We concentrate our research on the Current Account of the Balance of Payments. In this, the economic flows into and out of the country are classified into the broad categories: goods, services, profit/income flows and current transfers.

Balance of Payments data have been compiled by the Central Statistics Office on a quarterly basis since Quarter 1, 1981. There have been certain changes in the definitions of various series during the past 20 years, but we have been able to produce consistent time series for 15 main BOP components. Our data set consists of 75 observations (1981Q1–1999Q3) of this 15 dimensional vector time series.

The idea of co-integration was introduced by Granger in 1981. The principle of co-integration is simple: an equilibrium relationship among a set of non-stationary variables implies that their stochastic trends must be linked. The equilibrium relationship implies that the variables cannot move independently of each other.

Definition: Co-integrated *(Granger 1981, Engle and Granger 1987)*
The components of a vector time series $\mathbf{x_t}$ are said to be co-integrated of order d,b, denoted $\mathbf{x_t} \sim \mathbf{CI}(\mathbf{d}, \mathbf{b})$, if

1. all components of $\mathbf{x_t}$ are I(d) (Integrable of Order d);
2. there exists a vector $\alpha \neq \mathbf{0}$ so that $\mathbf{z_t} = \alpha \mathbf{x_t} \sim \mathbf{I}(\mathbf{d} - \mathbf{b}), \mathbf{d} \geq \mathbf{b} > \mathbf{0}$. The vector α is called the co-integrating vector.

Because of the dominance of Multinational Direct Investment Enterprises in Ireland, most of the growth in Ireland's trade surplus has been matched by a similar growth in service imports and profit outflows. It is of interest to find the equilibrium relationship among them.

Our analysis began by seasonally adjusting the data using the US Census Bureau's X-11 seasonal adjustment procedure. We then performed a standard co-integration analysis, looking for a set of series which are I(1) and for a linear combination of these I(1) series which is stationary.

This initial analysis failed because Augmented Dickey Fuller (ADF) tests indicate that over the period 1981Q1-1999Q3 many of the series do not appear to be I(1) (the first differenced series are not themselves stationary). If however we restrict our reference period to 1998Q1 1996Q2, then we find that there are some series that have unit roots according to the ADF tests. Using Johansen's Maximum

Likelihood Co-Integration procedure we also find that in this period there is possible co-integration between three of the main BOP components.

The next phase of our research involved looking for possible I(2) components in the data. These are series whose second differences are stationary. It appears that some of the individual series may indeed be I(2). We use Johansen's procedure for I(2) variables, consisting of two reduced rank regressions, to examine the possibility of Polynomial Co-Integration in the VAR model. Here we are looking for linear combinations of the original series and their first differences that are stationary. We are continuing to examine the I(2) model.

While our investigations are not complete we can say that the application of the technique of co-integration to Ireland's Balance of Payments has provided some insights and we hope that more will be forthcoming.

Most readers would be aware that many economic series would generally be I(1). The indication that some of Ireland's Balance of Payments series were I(2) and not I(1) may initially appear surprising, but it is just another effect of the ubiquitous "Celtic Tiger".

Estimating the Return from Genetic Enhancement in Milk Production

John A. Curtis
Economic and Social Research Institute, 4 Burlington Road, Dublin 4

Abstract

Introduction

Genetic enhancement is an age–old practice in crop and animal husbandry systems and long pre–dates the recent debate on genetically modified organisms. Crops and animals are bred to enhance their favourable traits and to diminish the presence of their less desirable characteristics. The purpose of genetic enhancement in the dairy industry is to improve efficiency in milk production and to produce a product that is best suited to market requirements, e.g., with respect to the fat and protein content of milk. It is well documented that genetic enhancement improves milk production (see for example, Bebber et al., 1997; Grosshans et al., 1997; Norman and Powell, 1999; Buckley et al., 2000a,b) but the important economic question is under what circumstance is investment in genetic enhancement profitable? This paper is concerned with the micro level decisions facing a farmer deciding on a breeding program. The farmer decides what sires to use based on their price, expected progeny performance and also the monetary return of improved output. The object of the paper is to measure the physical return on the genetic enhancement investment in a particular dairy herd and determine the circumstances under which the investment is profitable. Genetic enhancement is measured by indices of sire's genetic enhancement ability that are comparable across all sires. In the process of estimating the return from genetic enhancement the reliability of various genetic enhancement indices as tools to predict future milk production performance will be examined. Knowing the likely return from genetic enhancement and the reliability of various genetic indices has obvious benefits to farmers deciding a breeding programme.

To measure the return of genetic enhancement in milk production the paper proceeds in two stages. The first stage is to estimate the physical return from genetic enhancement investment. The estimate of the physical return is based on production statistics and the breeding program of a single dairy herd. The return on an investment in an Artificial Insemination sire is the value of the lifetime improved milk production from the sire's progeny. The second stage of the paper involves comparing the value of improved lifetime milk production performance to the cost of the initial investment in AI sires.

Estimating the Physical Return from Genetic Enhancement

Genetic Enhancement

In a way similar to racehorses going to stud farmers buy in the services of various sires depending on their requirements rather than rely on one or more on-farm bulls. The best dairy cows are usually bred with dairy herd replacement in mind and the choice of sire in this circumstance will depend on the expected milk production characteristics of the sire's progeny. The progeny's characteristics obviously depend on the characteristics inherited from both the dam and sire and to assist in the choice of sire the cattle breeding federations have standard procedures for comparing animals. In assessing a sire a farmer may look at two types of assessment, one for conformation and another for milk production though this analysis will only focus on milk production characteristics or proofs. The milk production proofs are

based on the milk production performance of at least 75 of the sires' daughters across at least 50 different herds. The proofs are expressed on the same base and scale and are directly comparable to each other.

Milk Production

The goal in modelling milk production is to adequately capture the biology involved. Milk production depends on environmental factors such as weather and housing, and on the quality and availability of fodder. Production also depends on a cow's age or lactation, when a cow calved and the number of days since calving. A cow's milk production usually increases in the first several lactations (Buckley et al., 2000b). The availability of fodder in a grass-based system has relevance for when a cow calves and its cycle of milk production. About 6–9 weeks after calving a cow's yield reaches its maximum before declining gradually to a dry period before calving (Buckley et al., 2000b). Environmental inputs, such as weather and fodder as well as health and management factors contribute to 65% of actual milk yield. Herd specific factors account for about 30% of the total variance in milk and fat yield in large multi-herd samples (Chauhan, 1987).

Methodology

The contribution of genetic enhancement to milk production will be estimated by means of a milk production function. The panel nature of the data would suggest that panel data econometrics should be used to estimate the production function but two difficulties arise if we wish to use standard panel data econometric models. The first relates to the dimension of the panel. The usual dimensions for a panel data set are cross section observations through time. The milk production data are cross section observations of cows but there are two dimensions of time, which are monthly during the milking season and three milking seasons. The second time dimension is required to control for the seasonal nature of production. The second difficulty in modelling milk production is that there is not just one output produced. Depending on its ultimate use milk is valued for its composition of fat or protein and is now frequently sold based on protein content. Modelling production therefore entails estimating three production functions, preferably simultaneously.

The most desirable approach would be to estimate a system of production functions and recognise the panel nature of the data. To avoid the double time dimension problem either a single panel with annual production or yearly panels with monthly data could be estimated. With limited observations an annual production panel is not practical. Estimation therefore will be based on monthly data for each of the three years available. Systems estimators for panel data are not readily available and for this initial exercise a systems estimator for panel data is not developed. Instead both a systems estimator and a panel data estimator are used separately. We will see that the primary difference between the estimates of the two approaches is the estimates of the covariance matrix.

Consistent with the science literature examining genetic merit and performance a linear production function is assumed for both estimators (Buckley *et al.*,2000; Ibáñez *et al.*, 1999; Olori *et al.*, 1999). The specification of the production function will differ depending on the estimator used. To estimate a system of production functions we assume that there is only one endogenous variable per production function in which case we can use the seemingly unrelated regressions estimator. The m'th equation of the production system is specified as $y_m = x_m\beta_m + \varepsilon_m$. Where $y_m(T \times 1)$ represents the m'th milk output, with T being the number of monthly milk recordings in the year. The explanatory variables are included in x_m. The residual ε_m is assumed to be distributed normally with mean zero and variance-

covariance $\sigma_{mp}\mathbf{I}_T(m, p = 1, 2, 3)$, where $\mathbf{I}_T$ is an identity matrix of order T.

For the panel data model each of the three production functions are separately specified as follows: $y_{it} = x_{it}\beta_1 + s_i\beta_2 + \nu_i + \varepsilon_{it}$, where y_{it} is cow i's output at time $t(t = 1..T)$. The units of output are measured as day-tests of protein, butterfat, and the remaining milk and the time periods refer to specific days in the milking season during a single year. Variables that vary over time and cows are included in x_{it} and consist of number of days milking and somatic cell count measure. Variable s_i incorporates genetic variables, lactation and calving date that do not change over time. There are no time-only varying variables included in model estimation. In the panel model $\nu_i + \varepsilon_i t$ is the residual error, one component of which is time invariant. The assumptions on the error term are that $E(\nu_i) = E(\varepsilon_i t) = 0, E(\nu_i\varepsilon_{it}) = 0, E(\varepsilon_{it}\varepsilon_{js}) = \sigma_\varepsilon^2$ for i = j, t = s and 0 otherwise; $E(\nu_i\nu_j) = \sigma_\nu^2$ for i = j and =0 for i $\neq$ j. Both components of the error are also assumed to be uncorrelated with either $x_i t$ or s_i. Our particular interest is the estimate of β_2, which includes the parameters that measure the effect of genetic merit on milk production.

References

Bebber, J. van, Reinsch,N., Junge,W., Kalm, E., 1997. Accounting for Herd, Year and Season Effects in Genetic Evaluations of Dairy Cattle: A Review. *Livestock Production Science* 51: 191 – 203.

Buckley, F., Dillon, P., Crosse, S., Flynn, F., Rath, M. 2000b. The Performance of Holstein Friesan Dairy Cows of High and Medium Genetic Merit for Milk Production on Grass-Based Feeding Systems. *Livestock Production Science* 64: 107 – 19.

Buckley, F., Dillon, P., Rath, M., Veerkamp, R.F. 2000a. The Relationship Between Genetic Merit for Yield and Live Weight, Condition Score, and Energy Balance of Spring Calving Holstein Friesian Dairy Cows on Grass Based Systems of Milk Production. *Journal of Dairy Science* 83, no. 8: 1878–86.

Chauhan, V.P.S. 1987. Partitioning of Herd, Year and Season Variation in Milk Production. *Livestock Production Science* 16: 107–16.

Grosshans, T., Xu, Z.Z., Burton, L.J., Johnson, D.L., Macmillan, K.L. 1997. Performance and Genetic Parameters for Fertility of Seasonal Dairy Cows in New Zealand. *Livestock Production Science* 51: 41–51.

Ibáñez, M.A., Carabañ o, M.J., Alenda, R. 1999. Identification of Sources of Heterogeneous Residual and Genetic Variances in Milk Yield Data from Spanish Holstein–Friesian Population and Impact on Genetic Evaluation. *Livestock Production Science* 59: 33–49.

Norman, H.D., Powell, R.L. 1999. Dairy Cows of High Genetic Merit for Yields of Milk, Fat and Protein. *Asian–Australian Journal of Animal Science* 12, no. 8: 1316–23.

Olori, V.E., Hill, W.G., McGuirk, B.J., Brotherstone, S. 1999. Estimating Variance Components for Test Day Milk Records by Restricted Maximum Likelihood with a Random Regression Animal Model. *Livestock Production Science* 61, no. 53–63.

A Graphical Model Approach to Complex Problems in Genetics

Nuala A. Sheehan
Department of Epidemiology and Public Health
University of Leicester UK

Abstract

Probability and likelihood computations are essential in any analysis of genetic data on groups of related individuals or *pedigrees*. These calculations are relevant to applications in such diverse areas as, genetic counselling, selective animal breeding, inference on the genetic nature of a disease, analysis of surviving genes in an endangered species and linkage analysis. An exact method for computing probabilities on pedigrees, in which at least one of every parent pair is a founder, was proposed by Elston and Stewart ((1971) Hum.Hered.**21:**523–542) and was finally generalised by Cannings, Thompson and Skolnick ((1978) Adv.Appl.Prob.**10:**26–61) to include arbitrarily complex pedigrees and genetic models. This method has become known in the statistical genetics literature as *peeling* and is essentially the same idea as that independently developed ten years later in an expert systems context (Lauritzen and Spiegelhalter (1988) J R Stat Soc B **50:**157–224) for the calculation of posterior probabilities on general Bayesian networks. Although in theory, any pedigree can be peeled, in practice, exact methods break down completely when the pedigree has too many interconnecting cycles or *loops*. This is because of the enormous storage requirements involved. The loops which make a pedigree complex for computational purposes are typically caused by inbreeding relationships or multiple inter-marital relationships.

The ever-increasing numbers of polymorphic markers currently being made available (Sobel and Lange (1996) Am J Hum Genet **58:**1323–1337) are causing the computational problems to intensify in genetic applications. This is particularly apparent in the area of linkage analysis, for example, where joint probabilities over large numbers of genetic loci are required. For almost any analysis, peeling fails on the large complex pedigrees which frequently arise in animal populations. As a consequence, pedigree information is either discarded completely and exact analyses performed on simple substructures, or the structure is approximated by one which retains as many features as possible but which can still be peeled (Wang at al (1996) Theor.Appl.Genet.**93:** 1299–1309). Alternatively, Markov chain Monte Carlo (MCMC) methods (Hastings (1970) Biometrika **57**(1):97–109) can be employed to provide estimates of probabilities and likelihoods of interest on the true structure (Thompson (1994) Stat.Sci. **9** (3) 355-366 for, example). However, MCMC methods have not really been tested extensively on these large problems and tend to be viewed with some suspicion, in practice, due to the unreliability of the resulting estimates (Hoeschele et al (1997) Genetics **147:** 1445–1457).

The problem to be discussed here is that of detecting a quantitative trait locus (QTL) from possibly incomplete marker data on general pedigrees. The particular approach to be taken involves the use of graphical models. An argument that graphical models provide the ideal framework for the development and testing of different MCMC sampling schemes, crucial to any real progress in this area, is also presented. Although graphical models feature explicitly in several specific applications in genetics (Kong (1991) Genet Epid **8:** 81–103, Jensen and Kong (1999) Am J Hum Genet **65:** 885–901, Lund and

Jensen (1999) Genet Sel Evol **31:** 3–24), the general applicability of this approach to solving complex problems in genetics has not been widely appreciated. Yet, the natural modularity inherent in these problems makes them ideal for such a representation. The use of graphs in genetics dates back to the path analysis diagrams of Wright ((1934) Ann.Math.Stat.**5:** 161–215). Indeed the standard representation of a pedigree as a marriage node graph is itself a graphical model representing the qualitative aspects of Mendelian inheritance by which an individual's genetic properties depend only on the genes of his parents (Spiegelhalter (1998) Appl.Stat. **47** (1) 115–133). The graphical model approach aims to fully exploit *all* the conditional independence structures of the problem at hand and perform calculations at the most local level possible. Furthermore, the flexibility of these models makes them readily adaptable to any changes in the pedigree or the genetic model.

The aim of this talk is to attempt to dispel some of the suspicion surrounding graphical models by explaining what they are and to demonstrate their relevance to computational problems in genetics. The ideas will be illustrated with an application to a simple QTL-mapping problem on a half-sib design.

Parallel Algorithms for Bayesian Inference of Spatial Gaussian Models

Matt Whiley
Trinity College Dublin

Abstract

The advantage of parallel processing over serial processing is that it allows us to consider problems that would otherwise require too much storage or take too long to compute.

For statisticians a common computationally intensive task is Markov chain Monte Carlo (MCMC). Here we consider one particular MCMC application, that of fitting spatial Gaussian models, and consider the effectiveness of using parallel algorithms in this example.

Spatial Gaussian models are widely used to model spatial data. We have a set of locations $\{s_1, s_2, \ldots, s_N\}$ and a process $\{Y_1, Y_2, \ldots, Y_n\}$ defined at these locations. This process is assumed to have a Gaussian distribution with zero mean and we further assume that the correlation between the process at locations s_i and s_j is a function of the distance, $d(s_i, s_j)$, and parameters θ. A typical example of this might be $\rho(Y_i, Y_j) = e^{-3d(s_i,s_j)/\theta}$.

Here we consider statistical inference under a particular generalisation of this model as described in Diggle *et. al* (1998). Poisson random variables $X_1, X_2, \ldots, X_n$, defined at locations $s_1, s_2, \ldots, s_n$ are assumed to be conditionally independent given $Y_1, Y_2, \ldots, Y_n$. The dependence of X_i on Y_i is through the conditional mean $\mathbb{E}(X_i \mid Y_i)$, taking the form $\log(\mathbb{E}(X_i \mid Y_i)) = Y_i + \beta$.

It is natural to parameterise the correlation structure in such a problem. However this introduces computational difficulties when calculating the likelihood, which is a function of the inverse of the covariance matrix. This computational difficulty is both in terms of the storage space required for the covariance matrix and the computational expense in determining its inverse. This latter problem is of particular relevance to MCMC techniques in which the likelihood must be evaluated at very many different parameter values, each involving the inversion of a different covariance matrix.

We consider the application of parallel processing to the computation of these matrix inverses. Our conclusion is that although some speed up is obtained it is not as much as might be hoped. Reasons for this are given and possible alternative parallelisations of these MCMC algorithms are discussed.

References

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998), Model-based geostatistics (with discussion), *Applied Statistics, vol. 47, no. 3, pp. 299–350.*

Modelling Uncertainty in Fatigue Criteria

Simon P. Wilson
Department of Statistics, Trinity College Dublin

Abstract

The objective of a multi-axial fatigue criteria is to determine, from given material properties and a specification of the full stress tensor acting on a component, whether the component is subject to stresses that will cause fatigue failure. Many such criteria have been proposed, starting in the 1950's; papers by Papadopoulos (1987) or Ballard (1995) survey the literature. In this talk, we describe a method to quantify the uncertainty in a criterion value that arises from random variability in material properties and experimental error. We do not consider the case of random loading. Our approach is to estimate the uncertainty by Bayesian statistical methods, implemented with Monte Carlo simulation, and is illustrated with examples.

1. Fatigue Failure under Simple Loadings

In a simple loading experiment, a specimen is subjected to a periodic bending stress that bends it backwards and forwards in one direction. An alternative is a simple twisting motion. Four things are observed about these experiments on metal components:

1. Specimens will eventually fracture at stresses far below the stress that would break them;
2. The fracture is caused by the initiation and propagation of cracks in the specimen;
3. The number of stress cycles until such *fatigue failure* is a decreasing function of the maximum stress experienced;
4. In metals there is a maximum stress limit below which fatigue failure does not occur, or only occurs after a very large number of cycles.

The stress below which the specimen will not fail is called the *fatigue limit*. Clearly, if one is designing a component that is subject to such simple loadings, then one should ensure that the maximum loading is smaller than the fatigue limit. This leads to a simple fatigue criteria for a component, to determine if it will fail due to fatigue:

Fatigue failure will occur **if and only if** *Max. stress level* $<$ *Fatigue limit*

The fatigue limit is found by experiment. Several specimens are subjected to loadings of different amplitudes, the number of cycles to failure is recorded, a curve fitted to these data and the limit estimated from the curve. There is variation in the cycles to failure at each amplitude, because of random variation in material properties, leading to uncertainty in the estimate of the limit. Thus, there is uncertainty in whether the fatigue criteria has satisfied. In this talk, we discuss how to compute P(Fatigue failure will occur | data). We use a Bayesian approach to the fitting of the data and estimation of this probability, implemented by Monte Carlo simulation.

2. Multi–axial Loadings

Most real components do not experience such simple loadings. In fact, they are subject to multi–axial loadings at several different points, in different directions and motions, and of varying strengths and frequencies, leading to a very complicated pattern of stresses that vary as a function of space and time in the component. These stresses are described by a function known as the stress tensor $\sum(x,t)$, which is a function of location x in the component and time. The stress tensor is a vector quantity; it requires 6 values at each point and time to completely describe the stress.

Do the ideas of fatigue limit and a fatigue criteria extend to these more complicated situations? The answer is yes, but at the expense of considerably more mathematics. Several multi-axial fatigue criteria have been proposed. They all take the following general form:

$$\textit{Fatigue failure will occur } \textbf{if and only if } f(\textstyle\sum(x,t),\theta) < 0,$$

where θ represents material properties and f is some function. Each fatigue criteria proposed has a different f and may use different material properties. The functions f are usually complicated (examples will be given in the talk) but the material properties are usually just two simple fatigue limits: one in simple bending and one in simple twisting.

3. Uncertainty in Multi–axial Fatigue Criteria

So, multi-axial fatigue criteria are a function of simple fatigue limits, which we have observed are subject to uncertainty, being estimated from experimental data. So we come to the main objective of this work, which is how we can propagate the uncertainty that we have in estimates of the simple fatigue limit to produce an estimate for P(Fatigue failure will occur | data) in the multi-axial case. This just requires some uses of the laws of probability, implemented by Monte Carlo simulation. In the talk, we show how to do this and illustrate with an example from data on a cast iron.

4. Extensions of the Approach

To conclude the talk, we discuss how to extend the approach to the situation here loadings (and therefore $\sum(x,t)$) are random.

References

Ballard, O., Dang Van, K., Deperrois, A. and Papadopoulos, Y.V. (1995). High cycle fatigue and a finite element analysis, *Fatigue Fract \ Engng. Mater. Struct., Vol. 18, no. 3, pp. 397-411.*

Papadopoulos, Y.V. (1987). Fatigue polycyclique des mtaux: une nouvelle approche. Ph.D. thesis, cole Nationale des Ponts et Chausses, Paris.