ROYAL STATISTICAL SOCIETY

RSS 2013 International Conference

Northumbria University, Newcastle

2-5 September 2013

ABSTRACTS BOOKLET

Abstracts are ordered in session order for oral presentations followed by poster presentations

1.1 Contributed – Communicating Statistics

Tuesday 2 September 9.20am - 10.20am

EMOS – The European Master in Official Statistics

Markus Zwick
Eurostat, Luxembourg, Luxembourg

EMOS is an infrastructure project aimed at developing a program for Training and Education in Official Statistics. The idea is to set up conditions for labelling courses and university programmes, which would lead to a qualification in European Official Statistics. EMOS is a joint project of different stakeholders with the aim of reaching a higher level of knowledge in various ways:

- Firstly, statistical producers could benefit from young and well-qualified researchers in official statistics.
- Secondly, other organisations with a link to statistics (ministries, central banks, research institutes, consulters) could acquire better qualified staff in statistics on the labour market.
- A third point is that NSI and Universities stand to learn a great deal from each other by having this project in common.

In spring 2012, Eurostat launched a Call for Tender for the feasibility study 'Towards a European Master in Official Statistics'. The purpose of this study is to contribute to the creation of a European Master in Official Statistics and to create a network at a European level.

The feasibility study started in December 2012 and is expected to draw to a close after twelve months with a final technical report. Based on the feasibility study and assuming that the universities and relevant stakeholders remain interested and the systems of education across Europe are not too different, it is expected that the first courses of EMOS could start in autumn 2014.

The presentation should inform all interested stakeholders about the activities of Eurostat in relation to EMOS.

1.1 Contributed – Communicating Statistics

Tuesday 2 September 9.20am - 10.20am

Making Data Accessible – JRF DATA. The new data initiative of the Joseph Rowntree Foundation

Aleks Collingwood
Joseph Rowntree Foundation, York, UK

The objective of the presentation is to let the audience know about our brand new initiative and one of our major outputs this year, **JRF DATA**. The aim of this initiative is to make JRF the place to get the facts about Poverty, Place/housing and our Ageing Society. The information is presented and available to a wide audience in an accessible way. The process behind the development of this initiative will be explored along with how it can be used as an invaluable reference tool.

Our website has been developed to enable us to launch 100 'indicator pages' this year. The statistical indicators selected cover our three theme areas. Examples include child poverty and homelessness. Each page features a user friendly interactive graph/map, a paragraph that gives JRF's take and interpretation of what the data means, a download feature (users can take the data with them), an embed feature (users can embed our charts/graphs on their pages/in reports), and links to relevant work in each area.

The result is an accessible and reliable reference tool for data in the areas of Poverty, Place and an Ageing Society. The indicators across the theme areas are not always mutually exclusive and one of the main strengths of the initiative is how they are interrelated and visually represented. Official datasets used for the analysis are nationally representative and in many cases the information is available by region and displayed in interactive maps. JRF DATA will be used by a wide audience from key stakeholders and data users to academia and the media.

1.1 Contributed – Communicating Statistics

Tuesday 2 September 9.20am - 10.20am

Census for all: dissemination of the 2011 Census Microdata Products

Johanna Hutchinson, Paul Waruszynski
Office of National Statistics, Hampshire, UK

Objectives

The investment of time and resources in a national census can only be justified if the results are accessible to users and meet their needs. Using 2011 Census Microdata as an example, this paper will demonstrate some of the ways ONS are communicating results. Microdata (samples of unit record level data) products are intended to target users who wish to obtain a broad overview and more demanding users who require more detailed and specific information about a topic. In addition, the release of Microdata necessitates ensuring data confidentiality, whilst maintaining utility in sample size and breadth of variables.

Method/Models

This presentation discusses the Microdata products, detailing the impact of stakeholder feedback on product development and the importance of targeted supporting documentation/software to ensure maximum user engagement.

Results and Conclusions

Five Microdata products will be released, split into individual and household components. Their specification is targeted to specific user groups. The impact of this on the inclusion of variables and level of detail will be discussed. The method of dissemination for each product is designed to meet user needs, whilst protecting data confidentiality and the success of this will be examined. Finally, the need to ensure suitable, targeted supporting documentation to engage novice users is given as an illustration of good practice and an example of promoting widespread engagement with census statistics.

1.2 Contributed – Communicating Statistics

Tuesday 2 September 9.20am - 10.20am

Simultaneous confidence bands for a percentile line in linear regression with application to drug stability studies

Yang Han[1], Wei Liu[1], Frank Bretz[2], Fang Wan[1]
[1]*S3RI and School of Mathematics, University of Southampton, Southampton, UK,*
[2]*Novartis Pharma AG, Basel, Switzerland*

Objectives

Simultaneous confidence bands have been used to quantify unknown functions in various statistical problems. A common statistical problem is to make inference about a percentile line in linear regression. Construction of simultaneous confidence bands for a percentile line has been considered by several authors. But only conservative symmetric bands, which use critical constants over the whole covariate range (-∞, ∞), are available in the literature. The main purpose of this research is to construct simultaneous confidence bands for a percentile line over a given covariate interval which can be finite or infinite, and to compare the bands under the average band width criterion.

Methods

Methods given in this paper allow the construction of exact symmetric simultaneous confidence bands for a percentile line over a finite interval of the covariate x. Furthermore, we propose a method of constructing an asymmetric simultaneous confidence band corresponding to each given symmetric band. We illustrate the proposed methods with a real example on drug stability study.

Results and Conclusions

Comparison under the average band width criterion shows that the exact symmetric bands can be substantially narrower than the corresponding conservative symmetric bands. Furthermore, we find that asymmetric confidence bands are uniformly, and can be very substantially, narrower than the corresponding exact symmetric bands. Therefore, asymmetric bands should always be used under the average band width criterion.

1.2 Contributed – Methods & Theory

Tuesday 2 September 9.20am - 10.20am

On the effects of the Diebold-Mariano test on the selection of prediction models

Robert Kunst[1,2], Mauro Costantini[3]
[1]Institute for Advanced Studies, Vienna (Wien), Austria, [2]University of Vienna, Vienna (Wien), Austria, [3]Brunel University, Uxbridge, UK

In evaluating prediction models, many researchers flank comparative ex-ante prediction experiments ('horse races') by significance tests on accuracy improvement, such as the Diebold-Mariano test. We argue that basing the choice of prediction models on such significance tests is problematic, as this practice tends to favour the null model, usually a simple benchmark. We explore the validity of the argument and quantify the effects by extensive Monte Carlo simulations with linear (ARMA) and nonlinear (SETAR) generating processes, for both nested and non-nested situations. In nested designs, the null distribution of the Diebold-Mariano statistic is accessed by the bootstrap-after-bootstrap method. The strength of the bias in favour of the null model varies across simulation designs. Generally, we find that utilization of additional significance tests in the selection of the forecast model fails to improve predictive accuracy relative to the decision suggested by a horse race comparison without any flanking significance testing.

1.2 Contributed – Methods & Theory

Tuesday 2 September 9.20am - 10.20am

Joint spatio-temporal modelling of bovine TB in badgers and cattle

Gabrielle Kelly
University College Dublin, Dublin, Ireland

The objectives of this study are to test if there is association between spatial associations of bovine TB (bTB) in cattle herds and badger (Meles meles Linnaeus) setts and to outline how linear geostatistical models (LGMs) and results relating to testing parameters on the boundary of hypothesis spaces may be used to do so.

Cattle herd and badger sett bTB incidence data are drawn from the Four Area Project, a formal badger removal project undertaken in four counties in Ireland from September 1997 to August 2002, to assess the effect of badger culling on the incidence of bTB. As important covariates differ for the two species, LGMs are fitted separately to data from each and the residuals combined. Sequences of LGMs are then fitted to the combined data and hypotheses related to spatial correlation structure are tested using critical values from mixtures of $\chi 2$ random variables.

Association was found between the spatial distribution of the disease in cattle and that in badgers in two of three areas and may be interpreted as evidence of cross-infection between the species. Separately, it was found spatial association of bTB in badger setts varies over time, between areas and with direction within an area. Similar results were found for cattle herds in agreement with previous analyses in the literature.

1.3 Contributed – Time Series

Tuesday 2 September 9.20am - 10.20am

Change-point detection of non-stationary time series using Wild Binary Segmentation

Karolos Korkas, Piotr Fryzlewicz
London School of Economics, London, UK

We propose a new technique for consistent estimation of the breakpoints of a linear time series where the number and locations are unknown using the Wild Binary Segmentation (WBS) of Fryzlewicz (2012). We adopt the nonparametric Locally Stationary Wavelet model which provides a description of the second-order structure of a piecewise-stationary process through wavelet periodograms estimated at multiple scales and locations. The advantage of WBS is its localisation feature which means that it works in cases where the spacings between breakpoints are very short. In addition, we improve the performance of the algorithm by combining the CUSUM statistics obtained at different scales and by using a post-processing step to eliminate spurious breakpoints. We provide an extensive simulation study to examine the size and power of our method for different types of scenarios.

1.3 Contributed – Time Series

Tuesday 2 September 9.20am - 10.20am

Quasi-maximum likelihood estimation of periodic autoregressive, conditionally heteroscedastic time series

Wolfgang Schmid, Florian Ziel
European University Viadrina, Frankfurt/O, Germany

We consider the general periodically stationary and ergodic causal time series model $Y_t = f(Y_{t-1},Y_{t-2},\ldots) + M(Y_{t-1},Y_{t-2},\ldots)Z_t$ with iid innovations Z_t. It nests various popular ones, such as the periodic ARMA-GARCH model. The asymptotics of the quasi-maximum likelihood (QML) estimation is discussed in detail. So we prove the existence of a solution, the consistency and asymptotic normality of the corresponding QML estimator under mild conditions. Applications to the multivariate nonlinear periodic AR(∞)-ARCH(∞) process with the periodic AR(∞)-APARCH(∞) and periodic ARFIMA-FIAPARCH models as special cases are shown. Due to the flexibility of this model we discuss some modelling issues. In detail we analyse Fourier and periodic B-spline approximation techniques for the periodic parameters. Furthermore we present applications to the hourly EUR/USD exchange rate time series and daily wind speed data.

1.3 Contributed – Time Series

Tuesday 2 September 9.20am - 10.20am

Big Data impacts on stochastic forecast models: evidence from FX time series

Sebastian Dietz
University of Passau, Passau, Germany

With the rise of the Big Data paradigm new tasks for prediction models appeared. In addition to the volume problem of such data sets nonlinearity becomes important, as the more detailed data sets contain also more comprehensive information, e.g. about non-regular seasonal or cyclical movements as well as jumps in time series. This essay compares two nonlinear methods for predicting a high frequency time series, the USD/Euro exchange rate. The first method investigated is Autoregressive Neural Network Processes (ARNN), a neural network based nonlinear extension of classical autoregressive process models from time series analysis. Its advantage is its simple but scalable time series process model architecture, which is able to include all kinds of nonlinearities based on the universal approximation theorem of Hornik, Stinchcombe and White 1989. However, restrictions related to the numeric estimation procedures limit the flexibility of the model. The alternative is a Support Vector Machine Model. The two methods compared have different approaches of error minimisation (Empirical error minimisation at the ARNN vs. structural error minimisation at the SVM). Our new finding is, that time series data classified as "Big Data" need new methods for statistical prediction. Estimation and prediction was performed using the statistical programming language R. Besides prediction results we will also discuss the impact of Big Data on data preparation and model validation steps.

1.4 Contributed – Healthcare

Tuesday 2 September 9.20am - 10.20am

Applications of mixed models to investigate progression of chronic diseases using routinely collected General Practice data: a case study in Chronic Kidney disease (CKD) in the UK

Zalihe Yarkiner[1], Rosie O'Neil[1], Gordon Hunter[1], Penelope Bidgood[1], Simon De Lusignan[2]
[1]Kingston Univeristy, London, UK, [2]University of Surrey, Guildford, UK

The development of new techniques and adaptation of existing methodologies for investigating large, complex, longitudinal data sets continues to be an important and developing area in the field of statistical modelling. General Practice (GP) patient records provide such a data source and have the potential to further knowledge and understanding of many aspects of public health service and provision.

Objectives

The aim of this research is to develop a longitudinal modelling framework for identifying factors related to the diagnosis and progression of chronic diseases using routinely collected GP records. The application is to Chronic Kidney Disease (CKD).

Methods

Mixed models are used for the analysis of repeated measures of kidney function found within individual patient histories. Linear mixed models provide an insight into the variability of disease progression both within and between patients, while accommodating the non-uniformity of time intervals between observations. However not all patients experience the same rate and pattern of progression and so non-linear mixed models are also employed in order to investigate the variability in rates of decline within the patient group.

Results

The results of our models, based on a sample of approximately 50,000 patients, have revealed much variation in rates of progression between patients, some of which can be explained by differences in the incidences of co-morbidities within patients but much of the variation is as yet unexplained. Current research is aimed at identifying causal patterns within, and the impacts of other associated factors on, rates of decline of CKD.

1.4 Contributed – Healthcare

Tuesday 2 September 9.20am - 10.20am

Using biological sample data to refine longitudinal measures of smoking behaviour

Lea Trela-Larsen, Jon Heron, Marcus Munafo
University of Bristol, Bristol, UK

We analysed repeated self-report measures of smoking, in a sample of 5,335 adolescents from 13 to 18 in a UK based birth cohort. Using Mplus we fitted a latent class cubic growth model, with varying individual age, to identify groups with differing trajectories of smoking behaviour in adolescence.

Individuals included in the analysis had smoking self-report data available at three or more of the six possible observation time points. The number of latent classes was chosen based on goodness of fit; Bayesian information criterion; Vuong-Lo-Mendell-Rubin likelihood ratio test; and bootstrap likelihood ratio test results.

Self-report data on smoking behaviour may exhibit bias; therefore biological measures can be useful to validate self-report data. Measures of cotinine, the principal metabolite of nicotine, were collected at age 15 from blood plasma. This presentation demonstrates how these biological sample data can be used to refine our latent class growth analysis, using the uncertainty around self-report to adjust the latent class trajectories.

1.4 Contributed – Healthcare

Tuesday 2 September 9.20am - 10.20am

Dealing with non-response in longitudinal analysis of trajectories of frailty and wellbeing in older people

Alan Marshall, Gindo Tampubolon
University of Manchester, Manchester, UK

This paper uses data from the English Longitudinal Study of Ageing (ELSA) to jointly model trajectories of health outcomes (frailty and wellbeing) and attrition thus accounting for the influence of missing data on model results. We illustrate that the influence of missingness varies according to the health outcome under investigation and develop a robust modelling approach to account for the effects of missing data in longitudinal analysis of health at the older ages.

Many longitudinal studies of the elderly are affected by the presence of missing values due to participants dying or dropping out of the study. For example, in ELSA the core sample falls from to 11,391 to 6,242 between waves 1 and 5. The patterns of missingness are not at random, for example, those who are frail are more likely to drop out of the study, so joint models of our variable of interest and the missingness process are needed in order to derive valid likelihood inferences. In this paper we apply a shared parameter model where two sets of equations (the missingness equation and substantive equation) are estimated simultaneously by specifying their full likelihood. We discuss why the modelling approach and results vary across measures of frailty and wellbeing and the biases of ignoring non-response in such settings.

1.5 Contributed – Heterogeneity

Tuesday 2 September 9.20am - 10.20am

Modelling heterogeneous variance-covariance components in two-level multilevel models with application to school effects educational research

George Leckie
University of Bristol, Bristol, UK

Applications of multilevel models – also known as hierarchical linear models, mixed-effects models, or random-coefficient models – to two-level continuous data nearly always assume a constant residual error variance at level-1 and constant random-effects variances and covariances at level-2. However, there is no reason why these homogeneity assumptions should hold in practice and in many educational and other applied studies it will be intrinsically interesting to relax them. In this paper, we extend the general two-level random-coefficient multilevel model by modelling the level-1 residual error variance as a function of predictors measured at both levels and we allow random-intercepts and random-coefficients to be included in this function. We model the level-2 variances and covariances as function of the level-2 predictors. We demonstrate, through simulation, that ignoring level-2 random effects in the level-1 variance function will estimate the level-1 variance function regression coefficients with spurious precision. We illustrate our approach through a step-by-step real data application to modelling school effects on student achievement. We fit our models using Markov chain Monte Carlo methods as implemented in the ESRC funded Stat-JR package under development at the Centre for Multilevel Modelling, University of Bristol.

1.5 Contributed – Heterogeneity

Tuesday 2 September 9.20am - 10.20am

Statistical inference on networks with heterogeneous degrees

Pierre-André Maugis, Sofia Olhede, Patrick Wolfe
University College London, London, UK

Observed networks often have nodes of heterogeneous degrees. One classical example is that of the repeatedly observed "exponential law" in social networks; a more concrete example being that of telephone networks: regular users only make few connections, while call centres make a large number of them.

Heterogeneousness makes statistical inference harder. The intuition of this being that it is easier to aggregate information among similar objects rather than different ones. Consider for instance the problems associated with high variance and heteroscedasticity in other fields of statistics. We aim here to address these problems in the case of networks' degree. To achieve this we will consider two cases: one where each node has a given expected degree and one where the said expected degree of each node is drawn independently from the same distribution. The first case allows us to describe the quality of each node separately (micro-analysis), while the second allows us to describe the characteristics of the population as a whole (macro analysis).

From a high level perspective, our results show that heterogeneousness among nodes translates into biased or more slowly converging estimation. However, the careful statistical analysis we performed enables us to remove the bias, or reduce it by several orders of magnitude, at little variance costs.

Finally we will show how the proposed models can be estimated using other motifs, or even collection of motifs – with a special focus on triangles – and relate this approach to the full maximum likelihood approach and the least square estimator.

1.5 Contributed – Heterogeneity

Tuesday 2 September 9.20am - 10.20am

A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses

Evan Kontopantelis, David Springate, David Reeves
University of Manchester, Manchester, UK

Heterogeneity has a key role in meta-analysis methods and can greatly affect conclusions. However, true levels of heterogeneity are unknown while researchers often assume homogeneity. We aim to: a) investigate the prevalence of unobserved heterogeneity and the validity of the assumption of homogeneity; b) assess the performance of various meta-analysis methods; c) apply the findings to published meta-analyses.

We accessed 57,397 meta-analyses from the Cochrane Library in August 2012. Using simulated data we assessed the performance of various meta-analysis methods in different scenarios. The prevalence of a zero heterogeneity estimate in the simulated scenarios was compared with that in the Cochrane data, to estimate the degree of unobserved heterogeneity in the latter. We re-analysed all meta-analyses using all methods and assessed the sensitivity of the statistical conclusions.

Levels of unobserved heterogeneity in the Cochrane data appeared to be high, especially for small meta-analyses. A bootstrapped version of the standard DerSimonian-Laird approach performed better both in detecting heterogeneity and providing more accurate overall effect estimates. Re-analysing all meta-analyses with this method we found that 17-20% of the statistical conclusions changed, when heterogeneity was detected with the standard model and ignored. Rates were much lower when the standard method did not detect heterogeneity or took it into account, between 1% and 3%.

Current practice assumes a zero between-study variance estimate leads to a more reliable meta-analysis. We found that a homogeneity assumption often results in a misleading analysis, since heterogeneity is very likely present but undetected. Finally, we must caution researchers against ignoring detected heterogeneity.

2.1 Invited – Communicating and interpreting statistical evidence in the administration of criminal justice

Tuesday 3 September 11.50am - 1.10pm

Statistics and the evaluation of evidence

Colin Aitken
University of Edinburgh, Edinburgh, UK

A necessarily brief review of the history of statistics in the evaluation of evidence will be given. It will start with the use of relative frequencies. A two-stage process was then introduced which assessed similarity in evidence from a crime scene and from a suspect by means of a significance test; if the results were not significant then a second stage in which rarity was assessed was conducted. Since the late 1970s, following a seminal paper by Dennis Lindley in Biometrika (1977), likelihood ratios have been adopted by many forensic scientists for evaluation of evidence. A likelihood ratio provides a measure which incorporates both similarity and rarity.

The likelihood ratio is not universally accepted as a panacea for the difficulties in evaluating evidence or for interpreting the value in a court of law. There is still much debate on topics such as relevance, probative value, inference, so-called 'Bayesianism' and the foundations of evidence (e.g. International Commentary on Evidence, 2010) which will be mentioned, again briefly.

2.1 Invited – Communicating and interpreting statistical evidence in the administration of criminal justice

Tuesday 3 September 11.50am - 1.10pm

Statistics, probability and the logic of forensic proof

Paul Roberts
University of Nottingham, Nottingham, UK

Statistical evidence, of one kind or another, is an increasingly familiar feature of modern criminal trials. The use of statistical information and probabilistic reasoning in litigation is often routine and uneventful. Occasionally, however, something apparently goes wrong and statistical evidence hits the headlines for all the wrong reasons, as in the tragic *Sally Clark* case. Or disputed issues of probability create a stir in more arcane professional circles, most notably in forensic scientists' least favourite Court of Appeal decision of recent memory, the footwear mark case *R* v *T* (2011). Indeed, these seemingly intermittent aberrations may prompt us to question whether all is truly well in the more routine and uneventful cases, in which statistical information and/or probabilistic reasoning may be playing an implicit and largely unexamined role. (Fingerprint examiners are one professional group who are slowly coming to terms with this dawning realisation.)

I am an academic criminal lawyer and legal theorist. Over the last five years or so, I have been collaborating closely with forensic statisticians (Colin Aitken, Roberto Puch-Solis) and experienced forensic scientists (Graham Jackson, Sue Pope) in the production of "practitioner manuals" which aim to provide reliable, effective, and above all intelligible advice to lawyers, judges, forensic scientists and other expert witnesses in the use of statistical evidence and probabilistic reasoning in criminal proceedings. I have learnt a great deal from my enormously knowledgeable, generous and patient collaborators (and plainly still have much more to learn). Perhaps the most important intellectual discovery that I have made in this time, however, is one that I deduced for myself, and which casts lawyers' alleged numerophobia in a new light: namely, that much of the difficulty surrounding statistical evidence and probabilistic reasoning has *absolutely nothing to do with statistical or probability theory, mathematical axioms, computation or anything of that kind*. Instead, it relates to a set of more fundamental, strictly philosophical, linguistic and legal issues pertaining to the logic of proof.

In this presentation I will share my "discovery" and reflect upon its significance, with practical illustrations drawn from interdisciplinary academic literature, recent case-law and current proposals for institutional reform.

2.1 Invited – Communicating and interpreting statistical evidence in the administration of criminal justice

Tuesday 3 September 11.50am - 1.10pm

The basis of opinions in forensic science

Graham Jackson
University of Abertay, Dundee, UK

The primary function of forensic science in criminal courts is the provision of expert opinion to help the court reach a decision on whether a defendant has or has not committed a stated offence. Perhaps surprisingly, there has been little formal training traditionally for forensic scientists and other experts on how to provide opinions that meet fundamental criteria of logic, balance, robustness and comprehensibility. There has always been an acknowledgment that the significance of a scientist's findings is related generally to an assessment of frequencies of occurrence but there has been no formal means of incorporating such data and expert knowledge logically and coherently. Over the last two decades, a formal structure, based on a Bayesian paradigm, has been gradually developed and is slowly being adopted, albeit with some reticence and difficulty, across the profession. The structure helps experts formalise the basis of their opinions and provides a means by which police, lawyers and courts may critique and challenge those opinions.

This presentation will explain and explore two key notions from this formal structure – the hierarchy of issues and the classification of opinions – and will illustrate how these may be applied to help provide supportable opinions and to inform decisions on the examination strategy to be employed in a case.

2.3 Invited – Data sharing and linking – the methodological, legal and practical issues

Tuesday 3 September 11.50am - 1.10pm

Multiple imputation for linking data

Katie Harron[1], Harvey Goldstein[2]
[1]UCL Institute of Child Health, London, UK, [2]University of Bristol, Bristol, UK

Probabilistic record linkage techniques assign match weights to one or more potential matches for those individual records that cannot be assigned 'unequivocal matches' across data files. Existing methods select the single record having the maximum weight provided this weight is higher than an assigned threshold. The talk will discuss the problems associated with such an approach, which ignores all information from matches with lower weights, and for some individuals assigns no match. This leads to inefficiency and may also lead to biases in subsequent analysis of the linked data. It is proposed that a multiple imputation framework is utilised for data that belong to records that cannot be matched unequivocally. In this way the information from all potential matches is transferred through to the analysis stage. This procedure allows for the propagation of matching uncertainty through a full modelling process that preserves the data structure. Results are presented from several simulation examples that show how the procedure leads to reduced bias.

2.3 Invited – Data sharing and linking – the methodological, legal and practical issues

Tuesday 3 September 11.50am - 1.10pm

Improving our evidence base through making better use of existing administrative data: the DWP/HMRC/MoJ data sharing experience.

Melissa Cox[1], Josephine Daniels[2], Samaira Iniesta-Martinez[1]
[1]Ministry of Justice, UK, [2]Department for Work and Pensions, UK

Having rich data on the relationship between benefits, employment and offending is a goal for both the Ministry of Justice (MoJ) and the Department for Work and Pensions (DWP) so that, for example, analysis on what works to reduce re-offending can be undertaken. However there are financial and logistical challenges to achieving this, such as new data collections/surveys being costly. This is specifically relevant for disadvantaged groups where the costs of achieving good and representative response rates are particularly high. We have been working to improve our evidence base through making better use of existing administrative data.

In 2010 we received full legal and ethical approval for a one-off data share of administrative data between the DWP, HMRC and MoJ. This has led to valuable evidence on re-offending and welfare dependency that has contributed to policy making within DWP and MoJ. Due to the success of this initial data share, earlier this year we gained approval for data to be shared on an annual basis. Although this is a data sharing success story, there have been a few hurdles along the way.

This session will focus on the data sharing journey. Officials from MoJ and DWP will discuss the background to the data sharing project, the issues and challenges involved in data sharing and key lessons learnt. Discussion will then focus on how analysis from the linked data has supported policy development, plans for future analyses, and the potential scope for the wider research and statistical community to use this data.

3.2 Invited – George Casella, his life and work

Tuesday 3 September 2.30pm - 3.50pm

Convergence analysis of the Gibbs sampler for Bayesian general linear mixed models with improper priors

Jim Hobert[1], Jorge Roman[2]
[1]University of Florida, Gainesville, FL, USA, [2]Vanderbilt University, Nashville, TN, USA

A popular default prior for the general linear mixed model is an improper prior that takes a product form with a flat prior on the regression parameter, and so-called power priors on each of the variance components. I will describe a convergence rate analysis of the Gibbs samplers associated with these Bayesian models. The main result is a simple, easily-checked sufficient condition for geometric ergodicity of the Gibbs Markov chain. This sufficient condition will be compared and contrasted with Hobert & Casella's (1994) sufficient condition for posterior propriety. (This is joint work with Jorge Roman.)

3.2 Invited – George Casella, his life and work

Tuesday 3 September 2.30pm - 3.50pm

Consistency issues in variable selection

Elias Moreno[1], F. Javier Giron[2]
[1]University of Granada, Granada, Spain, [2]University of Malaga, Malaga, Spain

Consistency in variable selection in regression when the number of regressors grows as the sample size grows is considered. Pairwise consistency and posterior model consistency are compared, and some clues about the priors on models and on model parameters for variable selection in complex models are obtained.

3.3. Invited – Statistical challenges in quantitative finance

Tuesday 3 September 2.30pm - 3.50pm

Modelling bubbles and crashes in housing and stock markets

John Fry
University of Sheffield, Sheffield, UK

Based on tools and techniques originating from statistical physics we discuss modelling bubbles and crashes in financial markets. Markets operate by balancing risk and return. Though this may sound trite when modelling this feature mathematically, we are led to surprisingly subtle results that link naturally to phase-transition phenomena in complex systems – thus elucidating an oft-cited analogy in the literature. Potential applications include modelling bubbles and exogenous/endogenous shocks in stock markets and have attracted some interest from policy makers. Further, these initial approaches can be extended to incorporate the development of elementary technical trading strategies. We apply a multivariate version of our model to English house prices and uncover some interesting results regarding contagion across different regions across different periods of time. Our approach is practically minded; our models can be easily calibrated to real data and can be shown to have some relevance to the ongoing Euro crisis. Potential implications for policy makers will also be discussed.

3.3. Invited – Statistical challenges in quantitative finance

Tuesday 3 September 2.30pm - 3.50pm

Conditional alphas and realised betas

Valentina Corradi, Walter Distaso, Marcelo Fernandes
University of Warwick, Coventry, UK

This paper proposes a two-step procedure to back out the conditional alpha of a given stock from high-frequency returns. We first estimate the realised factor loadings of the stock, and then retrieve the conditional alpha by estimating the conditional expectation of the stock return in excess over the realised risk premia. The estimation method is fully nonparametric in stark contrast with the literature on conditional alphas and betas. Apart from the methodological contribution, we employ NYSE data to determine the main drivers of conditional alphas as well as to track mispricing over time. In addition, we assess economic relevance of our conditional alpha estimates by means of a market-neutral trading strategy that longs stocks with positive alphas and shorts stocks with negative alphas. The preliminary results are very promising.

3.4 Contributed – Meta-analysis

Tuesday 3 September 2.30pm - 3.50pm

Meta-analysis of time-to-event outcomes from randomised trials using restricted mean survival time: application to individual participant data

Yinghui Wei, Patrick Royston, Jayne Tierney, Mahesh Parmar
MRC Clinical Trials Unit Hub for Trials Methodology, London, UK

Meta-analysis of time-to-event trial outcomes commonly uses the hazard ratio (HR) as the treatment effect measure. However, the proportional hazards (PH) assumption may be violated for some or all included trials. An alternative measure of the treatment effect is the between-arm difference in the restricted mean survival time (RMST). For a given arm, the RMST is the expected time-to-event up to *t** and may be estimated as the integrated survival function *S(t)* from the time origin to a chosen *t**. Consistent estimation of RMST difference does not require PH. When divided by *t**, RMST difference quantifies the treatment-associated change in the mean survival probability up to *t*.* We study the potential role of RMST as an alternative to the HR in individual participant data (IPD) meta-analysis. The methods are illustrated in application to two IPD meta-analyses in cancer. Results include the estimated 5-yr RMST in each of the treatment arms, estimated 5-yr difference in RMST between arms and test results for non-proportional hazards across trials. We provide plots of the difference and mean difference in RMST against *t** for varying *t**, to visualize how treatment effects vary with time. We conclude that RMST and RMST difference are useful outcome measures in meta-analysis of time-to-event outcomes because they emphasize the often-neglected time dimension and avoid the PH assumption. The RMST difference is intuitively interpretable as ‘mean life-time gained/lost' up to a clinically relevant time horizon and particularly helpful in situations when treatment effects (HRs) may change with time.

3.4 Contributed – Meta-analysis

Tuesday 3 September 2.30pm - 3.50pm

Using meta-analysis of phase II trials to inform potential phase III trial results

Danielle Burke[1,2], Lucinda Billingham[1,2], Alan Girling[2], Richard Riley[1,2]
[1]MRC Midland Hub for Trials Methodology Research, Birmingham, UK, [2]University of Birmingham, Birmingham, UK

Objectives
Pharmaceutical companies use Phase II trial results to make decisions about proceeding to Phase III. We will show how a meta-analysis of results from multiple Phase II trials is very informative toward this decision.

Methods
We consider a meta-analysis of nine randomised Phase II trials comparing the efficacy of two therapies for acute myocardial infarction. Results for four outcomes were collected: intracranial haemorrhage, stroke, reinfarction and total mortality. We apply multivariate meta-analysis methods, and use the obtained summary results to predict the treatment effect on the four outcomes in a future trial. The multivariate meta-analysis approach jointly synthesizes all outcomes together whilst accounting for their correlation, to allow appropriate joint inferences across two or more outcomes. Predictions are formed by calculating 95% prediction intervals that account for the between-trial heterogeneity and the uncertainty in summary results. The methods are applied and compared in both frequentist and Bayesian frameworks.

Results and Conclusions
The meta-analyses show that the new treatment is promising for most outcomes. The calculated prediction intervals contain the treatment effects that were seen in subsequent Phase III trials. These Phase III results were described as contradictory to the Phase II results, but the prediction intervals reveal this is not the case. Our example demonstrates that the future results of Phase III trials can be predicted using 95% prediction intervals derived from the results of a Phase II meta-analysis. The Bayesian framework naturally allows estimates of the probability that the treatment will be beneficial in a new trial.

3.4 Contributed – Meta-analysis

Tuesday 3 September 2.30pm - 3.50pm

Multivariate meta-analysis using individual participant data, with application to continuous, survival and surrogate outcomes

Richard Riley[1], Michael Wardle[1], Malcolm Price[1], Christina Yap[1], Jan Staessen[2], Francois Gueyffier[3], Jiguang Wang[4]
[1]University of Birmingham, Birmingham, UK, [2]University of Leuven, Leuven, Belgium, [3]Inserm, Lyon, France, [4]Shanghai Jiaotong University School of Medicine, Shanghai, China

Objectives

Multivariate meta-analysis jointly synthesises effect estimates for multiple outcomes and accounts for their correlation. For example, in randomised trials evaluating hypertension treatment there often are continuous and survival outcomes, including: systolic and diastolic blood pressure (BP), stroke, cardiovascular disease, death from cardiovascular disease, and all-cause death. We consider multivariate meta-analysis of hypertension trials using individual participant data (IPD), to jointly evaluate the treatment effects across all outcomes and examine whether BP is a surrogate for survival.

Methods

Ten trials with IPD are available. Within each trial, treatment effect estimates and their variances are estimated for each outcome, and bootstrapping used to estimate their within-study correlations. A multivariate meta-analysis model is then estimated using frequentist and Bayesian frameworks to obtain summary results, make joint predictions across outcomes, and estimate correlations of the true outcome effects between-studies.

Results

The multivariate approach produces improved estimation of summary treatment effects, and allows joint inferences about combinations of outcomes. The within-study correlation between the treatment effects is high for systolic and diastolic BP; however it is surprisingly low (<0.1) between BP and survival outcomes. There are even some negative between-study correlations between the true treatment effects on BP and survival. The predicted probability of reducing cardiovascular risk by 10% in a population (trial) where the BP reduction is >5mmHg is also low. Thus BP appears a poor surrogate for cardiovascular risk.

Conclusion

Multivariate meta-analysis of IPD enables a more complete synthesis of multiple correlated outcomes, including joint inferences across outcomes and evaluation of surrogacy.

3.4 Contributed – Meta analysis

Tuesday 3 September 2.30pm - 3.50pm

Estimating the power of a meta-analysis using individual participant data

Joie Ensor, Karla Hemming, Richard Riley
University of Birmingham, Birmingham, UK

Background

Individual Participant Data (IPD) meta-analysis is becoming increasingly common, especially as the raw data allows the estimation of treatment-covariate interactions. However, before commissioning an IPD meta-analysis, researchers and grant bodies should understand the statistical power of the approach. This is rarely done because it is non-trivial and depends on numerous factors. We propose using simulation methods to estimate the power of an IPD meta-analysis, in relation to both overall treatment effect and treatment-covariate interactions.

Methods

For continuous outcome data, IPD from randomised trials are simulated by defining the number of studies, study sizes, baseline and follow-up means and standard deviations, a continuous patient-level covariate (mean and standard deviation), and the amount of between-study variability. The simulated IPD is then analysed using a suitable one-stage IPD meta-analysis model, and the treatment and interaction coefficients calculated. Repeating this process many times allows the power of the IPD meta-analysis to be estimated.

Results

We illustrate the simulation program in STATA and demonstrate how it might inform the design/funding of a new IPD meta-analysis project; for example, by quantifying how many IPD studies and patients are required to detect a true interaction with 80% power. We then use the program to ascertain the power of existing IPD meta-analysis, and show that, in hindsight, some are underpowered to answer their question.

Conclusion

Our findings show that the power to detect a true effect is not guaranteed when using IPD. Our simulation program now allows users to quantify the potential power of their IPD meta-analyses.

4.2 Contributed – Methods & Theory

Tuesday 3 September 4.20pm - 5.20pm

The evaluation of evidence for auto-correlated data with an example relating to traces of cocaine on banknotes

Amy Wilson[1], Colin Aitken[1], Richard Sleeman[2], Jim Carter[2]
[1]The University of Edinburgh, Edinburgh, UK, [2]MSA Ltd., Bristol, UK

Much research in recent years for evidence evaluation in forensic science has focussed on methods for determining the likelihood ratio where the data have been generated by various random phenomena. The likelihood of the evidence is calculated under each of two propositions, that proposed by the prosecution and that proposed by the defence. The value of the evidence is given by the ratio of the likelihoods associated with these two propositions. One form of evidence evaluation is related to discrimination in which the problem is one of source identity. The two propositions are that the source is or is not associated with criminal activity. The aim of this research is to evaluate this likelihood ratio under two explanations, one an extension of the other, for the random phenomena by which the data have been generated. The first is when the evidence consists of continuous auto-correlated data. The second is when the observed data are also believed to be driven by an underlying latent Markov chain. Four models have been developed to take these attributes into account: an autoregressive model of order one; a hidden Markov model with autocorrelation of lag one; and a nonparametric model with two different bandwidth selection methods. Application of these methods is illustrated with an example where the data relate to traces of cocaine on banknotes. The likelihood ratios using these four models are calculated for these data, and the results compared. The research is supported by an EPSRC CASE award, voucher number 009002219.

4.2 Contributed – Methods & Theory

Tuesday 3 September 4.20pm - 5.20pm

Direct semi-parametric estimation of fixed effects panel data varying coefficient models

Juan Manuel Rodriguez-Poo, Alexandra Soberon Velez
University of Cantabria, Santander/Cantabria, Spain

In this paper we present a new technique to estimate varying coefficient models of unknown form in a panel data framework where individual effects are arbitrarily correlated with the explanatory variables in an unknown way. The estimator is based in first differences and then a local linear regression is applied to estimate the unknown coefficients. To avoid a non-negligible asymptotic bias, we need to introduce a higher dimensional kernel weight. This enables us to remove the bias at the price of enlarging the variance term and hence achieving a slower rate of convergence. To overcome this problem we propose a one-step back-fitting algorithm that enables the resulting estimator to achieve optimal rates of convergence for this type of problem. It exhibits also the so-called oracle efficiency property. We also obtain the asymptotic distribution. Since the estimation procedure depends on the choice of a bandwidth matrix, we also provide a method to compute this matrix empirically. Monte Carlo results indicate good performance of the estimator in finite samples.

4.2 Contributed – Methods & Theory

Tuesday 3 September 4.20pm - 5.20pm

Nonparametric predictive inference for ordered three-class ROC analysis with continuous measurements

Tahani Coolen-Maturi[1], Frank Coolen[2], Faiza Elkhafifi[3]
[1]Durham University Business School, Durham, UK, [2]Department of Mathematical Sciences, Durham University, Durham, UK, [3]Benghazi University, Benghazi, Libya

Receiver operating characteristic (ROC) curves are widely used to assess the performance of a binary classifier. ROC curves have been used in many fields such as signal detection, medicine, radiology, biometrics, machine learning, data mining and credit scoring. ROC surfaces (3D surfaces) are currently used to assess the performance of three-class classifiers. Classification of a given (future) observation to one of three classes is an important task in many decision making problems. We present the nonparametric predictive inference (NPI) approach to three-ordered class ROC analysis, including results on the volumes under the ROC surfaces and consideration of the choice of decision thresholds for the classification.

4.3 Contributed – Business & Finance

Tuesday 3 September 4.20pm - 5.20pm

Small and medium-sized enterprises and the financial crisis

Jake Ansell
University of Edinburgh, Edinburgh, UK

Since the financial crisis that developed in 2007 banks have been urged to be prudent in lending, but also urged to lend more to small and medium-sized Enterprises (SMEs) through governmental initiatives. This paper explores a large part of UK SMEs' records during the 'credit crunch' from 2007 to 2010. The data consists of performance measures, demographics and financial information. Previously cross-sectional logistic models were employed, but these did not allow for time series effects, random effects logit panel models have been used to study SMEs' behaviour through the period. Another disadvantage of cross-sectional analysis is that macroeconomic variables cannot be brought effectively into the analysis, as they only consider a single time period. The panel analysis therefore allows macroeconomic variables into the data analysis. This should inherently improve the model's explanatory power, but it also allows examination of the impact of changes within the economy to be considered and potentially facilitates forecasting. The macroeconomic variables considered are indicators of directions of the economy, financial markets and general economic conditions. Newly established and matured SMEs are contrasted in the analysis to see if there has been a differential effect. Measures of model prediction accuracy are given for comparison across the models employed.

4.3 Contributed – Business & Finance

Tuesday 3 September 4.20pm - 5.20pm

Ensemble models: theory and applications

Silvia Figini[1], Marika Vezzoli[2]
[1]University of Pavia, Pavia, Italy, [2]University of Brescia, Brescia, Italy

When many competing models are available for estimation, model averaging represents an alternative to model selection. Despite model averaging approaches having been present in statistics for many years, only recently are they starting to receive attention in applications especially in credit risk modelling (see Figini and Fantazzini, 2009). In this paper we investigate model averaging and ensemble learning in order to obtain a well-calibrated credit risk model in terms of predictive accuracy. We compare Bayesian (see Steel, 2011, and the references therein) and classical model averaging approaches, like Random Forest (Breiman, 2001), Boosting (Freud and Schapire, 1996), and CRAGGING (Vezzoli and Zuccolotto, 2011) with the final aim of improving the predictive performance of the models. In this contribution we show how ensemble models can be usefully employed to obtain a well-calibrated model, in terms of predictive accuracy, for credit risk problems. The out-of-sample results show that on the basis of our proposal it is possible to obtain a good model that predicts credit default events, on the basis of the estimated probability of default with respect to the single classical models proposed in the literature to model credit risk data. Empirical evidences are provided on a real financial data set provided by a Credit Rating Agency.

4.3 Contributed – Business & Finance

Tuesday 3 September 4.20pm - 5.20pm

The Inverse of Autocovariance Matrix method for a better space-time modelling

Utriweni Mukhaiyar, Udjianna Pasaribu
Institut Teknologi Bandung, Bandung, Indonesia

The Inverse of Autocovariance Matrix (IAcM) approach has been introduced by Mukhaiyar et al. (2012) for space-time modeling especially the Generalized Space-Time Autoregressive (STAR) model. This approach was executed by investigating the process stationarity. It is analytically proved that the IAcM approach is better than the previous approach, which is used for checking the stationarity condition of Vector AR model through the eigenvalues of parameters matrix. In this study, we compare both approaches for modelling the space-time series. We use the monthly tea production of some plantations in West-Java Indonesia since January 1992 to December 2010 as the data. We apply a Generalised STAR model for the data since we assume that the present production in a certain plantation is a linear combination of past productions in the same and neighbour's sites. In this case, the spatial dependency is represented by a weight matrix whose main diagonal is zero and the sum of each row is one. This matrix is determined by considering that the nearer sites will give the larger credit for the referred site. We obtain that the IAcM approach is effective in model selection stage, since it eliminates more non-stationary possible models than the eigenvalues of parameter matrix approach. It also implies to a more efficient modelling. Furthermore in this case, the IAcM method gives a simpler model which shows better forecast productions.

4.4 Contributed – Medical: Scanning & Surveillance

Tuesday 3 September 4.20pm - 5.20pm

Estimation of Seasonal Influenza Vaccine Effectiveness (SIVE) in Scotland

Kimberley Kavanagh[1], Colin Simpson[2], Naz Lone[2], Lewis Ritchie[4], Chris Robertson[1,3], Aziz Sheikh[2,5], Jim McMenamin[2]
[1]University of Strathclyde, Glasgow, UK, [2]University of Edinburgh, Edinburgh, UK, [3]Health Protection Scotland, Glasgow, UK, [4]University of Aberdeen, Aberdeen, UK, [5]University of Maastricht, Maastricht, The Netherlands

Objective
We aim to estimate the effectiveness of the seasonal trivalent inactivated influenza vaccine in Scotland over a nine-year period using a unique set of linked electronic databases. Seasonal influenza is responsible for an estimated global three to five million cases of severe illness and 250,000 to 500,000 deaths per year. It is uncertain as to what extent national vaccination programmes can prevent this morbidity and mortality.

Methods
We undertook a linkage of patient-level primary care, hospital and death certification in a nationally representative cohort, yielding 1,767,919 person-seasons of observations. Using covariate adjusted Cox proportional hazards we estimate the effectiveness of the influenza vaccination in preventing influenza-related consultations, hospitalisations and deaths. Propensity score models are used to adjust for allocation bias in vaccine uptake. The potential effect of unmeasured confounding on estimates of vaccine effect (VE) was considered as a sensitivity analysis.

Results
Analysis of the pooled data over nine seasons for all individuals gave VE=19.3% (95% CI: (9.5, 28.0)%) against influenza-like illness consultations; VE=10.9% (95% CI: (3.8, 10.9)%) against influenza-related hospitalisations and VE= 29.1% (22.0, 35.6)%) against influenza-related deaths. To remove the effect of vaccination (VE=0) for the influenza related deaths outcome would require the presence of an unmeasured confounder which increased the risk of outcome by a factor of 4, was prevalent in 20% of the unvaccinated and 5% of the vaccinated individuals.

Conclusion
Vaccination was associated with significant reductions in the occurrence of the influenza related outcomes measured and estimates were robust to possible unmeasured confounding.

4.4 Contributed – Medical: Scanning & Surveillance

Tuesday 3 September 4.20pm - 5.20pm

Causal effects of changes in brain structure on behavioural and cognitive measures accounting for age-effects

Simon White, Fiona Matthews
MRC Biostatistics Unit, Cambridge, UK

The CamCAN study constitutes one of the largest neuroimaging studies to date, with detailed structural scans linked to a range of behavioural and cognitive measures. Although there is an expected decline in cognitive function with age, what constitutes health ageing is not clear. We investigate hypothesised causal effects between cognition and structure, and how they may be inferred from CamCAN.

The CamCAN sample is an observational study of individuals with sampling stratified by age, from 20 to 90 years, and gender. The imaging data are three dimensional brain scans which were pre-processed using standard imaging techniques and linked with behavioural measures covering memory, speech, intelligence, as well as demographic and epidemiological data collected during the study.

We investigate methods to make appropriate inference from the data within a Bayesian causal framework. Assessing causal effects using observational data presents several challenges, not least the confounding with overall age-related decline within the population.

The need to appropriately account for the age-related confounding is demonstrated. Several structural features, voxel-wise or using regions of interest, are shown to be linked to behavioural measures.

4.4 Contributed – Medical: Scanning & Surveillance

Tuesday 3 September 4.20pm - 5.20pm

Correcting for rater bias in scores on a continuous scale, with application to breast density

Matthew Sperrin[1], Lawrence Bardwell[2], Jamie C Sergeant[1], Susan Astley[1], Iain Buchan[1]
[1]University of Manchester, Manchester, UK, [2]Lancaster University, Lancaster, UK

Existing literature on inter-rater reliability focuses on quantifying the disagreement between raters. We introduce a method to *correct* for inter-rater disagreement (or observer bias), where raters are assigning scores on a continuous scale. To do this, we propose a two-stage approach. In the first stage, we standardise the distributions of rater scores to account for each rater's subjective interpretation of the continuous scale. In the second stage, we correct for case-mix differences between raters by exploiting pairwise information where two raters have read the same entity on a case.

We illustrate the use of our procedure on clinicians' visual assessments of breast density (a risk factor for breast cancer). After applying our procedure, 229 out of 1,398 women who were originally classified as high density were re-classified as non-high density, and 382 out of 12,348 women were re-classified from non-high to high density. A simulation study also demonstrates good performance of the proposed method over a range of scenarios.

4.5 Contributed – Communicating Statistics

Tuesday 3 September 4.20pm - 5.20pm

Back to basics: helpful concepts for reasoning and arguing that you might not have learned in your statistics course

Ulrike Naumann
'Institute of Cancer Research, Sutton, UK

In order to solve problems in statistics we usually require more than the ability to use statistical methods or software – we need to use reasoning skills and common sense for the following tasks:

- to understand statistical problems
- to understand issues involved in the research process
- to devise a structured analysis approach
- to explain methods and findings
- to critically read research literature
- to support teaching in statistics
- to explain statistical methods to a non-statistical audience without jargon

However, in our training as statisticians we rarely have a specific ‘reasoning’ training.

The presentation is designed to be of general interest to all professional statisticians, especially those who work in applied statistics, in teaching, and who need to engage with the general public.

I will give an overview of concepts relevant to reasoning and arguing from philosophy (based on R.J. Fogelin, W. Sinnott-Armstrong: ‘Understanding Arguments’). Using practical examples, I will show how these concepts are relevant for explaining statistical ideas, critically reading research literature and how they can be relevant in statistics projects.

The talk will cover different purposes and functions of arguments. Firstly, I will explain how the intermingling of different functions can lead to unclarity and misperceptions. I will then explain key concepts for analysing arguments. Finally, I will give an overview of fallacies in arguments and summarise possible ways to refute an argument.

4.5 Contributed – Communicating Statistics

Tuesday 3 September 4.20pm - 5.20pm

The use of interactive eBooks for teaching Bayesian statistical modelling and missing data methods using the Stat-JR package

William Browne[1], Richard Parker[1], Chris Charlton[1], Danius Michaelides[2], Camille Szmaragd[1], Harvey Goldstein[1]
[1]University of Bristol, Bristol, UK, [2]University of Southampton, Southampton, UK

As part of an ESRC funded grant we have developed a new statistical software package, Stat-JR (http://www.bristol.ac.uk/cmm/software/statjr/index.html) with many novel features including interoperability with most of the commonly used statistical software packages. The package uses a web browser as a user-friendly interface but also has a novel eBook interface. Our interactive eBooks combine the best features of books and statistical software packages and can embed (interactive) statistical analyses within the text of a web document. Thus, as the reader reads the document they interact with the book, for example changing parameters in a model, and the package then performs the modelling for the new inputs and updates the book accordingly. In this talk we will introduce and demonstrate Stat-JR and its eBook interface and show some of its features. These will include displaying MCMC algorithms that are specific to the chosen model and dataset, linking to other packages to use the best of their features or simply compare the estimates across packages for the same model. We will use as examples a standard multilevel statistical model and an eBook for missing data that uses new functionality for performing multiple imputation in situations where responses are at different levels of a hierarchy and may be continuous, binary or categorical.

4.5 Contributed – Communicating Statistics

Tuesday 3 September 4.20pm - 5.20pm

Communicating interactions – an alternative graphical view

Neil Spencer
University of Hertfordshire, Hatfield, Hertfordshire, UK

The objective of this paper is to provide a methodology for the graphical presentation of interactions in complex models. It thus provides a means to help statisticians communicate the meaning of interactions to non-statisticians (e.g. fellow members of a study team).

The use of interaction plots to help understand the nature of the complex relationships is well established and they frequently include margins of error. For models without extraneous variables (variables present in the model but not in the interaction plot), these can help identify the nature of the interactions. However, for those with discrete extraneous variables, margins of error are dependent on the choice of which category to hold constant or the process used to average over the categories. This means that the nature of the interaction is not always readily identifiable, particularly if the interaction is between two discrete variables and some combinations of categories have larger/smaller effects than others.

This work involves producing an alternative interaction plot where, for discrete variables involved in interactions, the plot shows the effects associated with differences between their categories. This makes it easier to see which categories have larger/smaller effects and whether these are significantly different from those for other categories. Multidimensional scaling is used on matrices of coefficients representing differences between categories. Margins of error are shown with overlapping or non-overlapping markers. The plots give a clearer picture for the interpretation of the interaction effects and aid the non-statistician (and statistician!) when attempting to understand and describe the model.

4.7 Contributed – Sports & Gaming

Tuesday 3 September 4.20pm - 5.20pm

A comparison of sporting heroes: Bayesian modelling of Test match cricketers

Pete Philipson[1], Richard Boys[2]
[1]Northumbria University, Newcastle upon Tyne, UK, [2]Newcastle University, Newcastle upon Tyne, UK

In this work the contentious problem, and eternal bar room debate, of comparing sportsmen whose careers took place in different eras is addressed. The application here is to Test match cricket, encompassing both batsmen and bowlers using data from the first Test in 1877 onwards. Direct comparisons are compromised by changes to the game itself over time, whether this is due to an expanding talent pool, fundamental changes to rules and equipment or other factors. The overlapping natures of careers is exploited to form a bridge from past to present. The overall aim is to compare all players simultaneously, rather than just relative to their contemporaries.

An additive log-linear model that incorporates year-specific and age-specific components is used to allow the innate ability of an individual to be identified. Particular attention is paid to the form of the ageing function and a range of alternatives will be considered. An assessment is also made as to whether ability is increasing over time using decade-specific hierarchical models. A Bayesian approach is adopted and the posterior distribution for model parameters is determined by using Markov Chain Monte Carlo (MCMC) methods with random walks. We use this posterior distribution to construct a table of leading batsmen and bowlers via their predictive distributions.

4.7 Contributed – Sports & Gaming

Tuesday 3 September 4.20pm - 5.20pm

Comparison of machine learning and statistical models with application in rugby fitness tests

Matthew Revie[1], Kevin Wilson[1], Robert Holdsworth[2], Stuart Yule[2]
[1]University of Strathclyde, Glasgow, UK, [2]Glasgow Warriors Rugby Club, Glasgow, UK

This study was motivated by collaboration with Glasgow Warriors Rugby Club. Between 30th January 2012 and 17th April 2012, each Glasgow Warriors rugby player completed a questionnaire every day. Glasgow Warriors believe that they can assess the players' general physical and mental wellbeing through the questionnaire. Over the same period, players infrequently completed a vertical jump test, which Glasgow Warriors use to measure the fitness and fatigues levels of players. In all, 430 vertical jump tests were carried out on 41 players. Glasgow Warriors wanted to develop a model that allowed them to predict two measures of primary interest, i.e. peak distance and peak power, on occasions when no vertical jump test was completed.

The objective of this study was to explore how different modelling approaches could capture the complex non-linear relationships in the data. In particular, standard models were unable to capture the bias introduced by each player when assessing their subjective beliefs. Linear mixed effects models and Support Vector Machines (SVM) were used to capture the effect of the subjective ratings given by each individual player. The modelling approaches were evaluated in terms of complexity, modelling accuracy and applicability. Broadly, the linear mixed effects model consistently outperformed the SVM; however, further study is required to explore the full range of SVM kernels.

4.7 Contributed – Sports & Gaming

Tuesday 3 September 4.20pm - 5.20pm

Statistical analysis of player behaviour of online Freemium games leading to business models

Anusua Singh Roy[1,2], Mark Robinson[2], Tracey Warner[2]
[1]Edinburgh Napier University, Edinburgh, Midlothian, UK, [2]Games Analytics, Edinburgh, Midlothian, UK

In recent years there has been an expansion in free-to-play computer games where money is made from advertising and players paying for additional features. To be successful in this increasingly crowded market requires attracting to and retaining players in games and incentivising them to purchase items. This work is focussed on developing approaches to analyse in-game behaviour of players and use that knowledge to improve retention and monetization within these games.

Statistical techniques such as segmentation, profiling, logistic regression and survival analysis can be used to extract knowledge from histories of player activity. The database is huge containing thousands of players and millions of game events triggered by them – truly Big Data. This raises a challenge to the Statistician. A generic approach has been developed by understanding players' progression through the game, comparing engaged to non-engaged players or payers to non-payers, segmenting different playing styles and developing models. This provides information on how to form strategies to encourage players to invite others, to remain playing the game and to buy game products using real rather than virtual currency.

Traditional techniques such as K means clustering, logistic regression and survival analysis implemented in the R computing environment have been found to be effective. The findings from the statistical approaches have been converted into business models which have lead directly through business success by expanding market share and helping game designers improve player retention in their game and generate higher revenues.

5.1 Contributed – Medical/Clinical Trials

Wednesday 4 September 9am - 10.20am

A review of statistical methodology for recurrent events, with application to major trials in heart failure

Jennifer Rogers, Stuart Pocock
London School of Hygiene and Tropical Medicine, London, UK

Composite outcomes are frequently adopted as primary endpoints in clinical trials as they take account of both the fatal and non-fatal consequences of the disease under study and lead to higher event rates. Such analyses of time-to-first-event are suboptimal for a chronic disease such as heart failure, characterised by recurrent hospitalisations, as relevant information on repeat events is ignored.

We shall illustrate and compare various methods of analysing data on repeat hospitalisations, using data from major trials in heart failure. In addition to describing each method and its estimated treatment effect and statistical significance, we investigate the statistical power using bootstrapping techniques.

Recurrent heart failure hospitalisations were analysed using the Andersen-Gill, Poisson and Negative Binomial methods. Death was incorporated into analyses by treating it as an additional event in the recurrent event process, and by considering methods that jointly model hospitalisations and mortality. We used a parametric joint frailty model to analyse the recurrent heart failure hospitalisations and time to cardiovascular death simultaneously.

Our analyses show that methods that take account of repeat hospital admissions demonstrate a larger treatment benefit than the conventional time-to-first-event analysis, even when accounting for death. Inclusion of recurrent events also leads to a considerable gain in statistical power compared to the time-to-first-event even approach. It seems plausible that in future heart failure trials, treatment benefit would not be confined to first hospitalisations only and so recurrent events should be routinely incorporated.

5.1 Contributed – Medical/Clinical Trials

Wednesday 4 September 9am - 10.20am

Flexible joint modelling of longitudinal and time-to-event data: a semi-parametric regression approach with exact likelihood

Jessica Barrett, Li Su
MRC Biostatistics Unit, Cambridge, UK

In existing joint models for longitudinal and time-to-event data with shared random effects, simple random effects such as intercepts and time slopes are usually included in the time-to-event sub-model to relate the level and progression rate of the longitudinal outcome for each subject to the occurrence of subsequent event (e.g. survival or dropout). In practice individual trajectories may be non-linear and cannot be adequately characterised by linear models with simple random intercepts and slopes, making it difficult to detect the true underlying relationship between the longitudinal and time-to-event outcomes. To overcome this problem, we propose a new joint model for longitudinal and time-to-event data. Specifically, we use penalised splines with truncated linear bases to flexibly model the non-linear patterns in individual longitudinal trajectories. A discrete time-to-event model is specified such that the discrete time intervals also determine the location of knots for the penalised splines in the longitudinal model. The penalised-spline coefficients are interpreted as the intercept and slope of the individual longitudinal trajectory in the corresponding time interval, which then enter the time-to-event model as time-varying covariates. This model structure offers considerable flexibility in characterizing the association between the longitudinal trajectories and time-to-event outcome. We use exact likelihood methods proposed by Barrett et al. (2013) for maximum likelihood estimation. The proposed methods are illustrated with CD4 count data from an AIDS trial with informative dropouts.

Barrett,J., Diggle,P.J., Henderson,R., Taylor-Robinson,D. Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference (submitted, 2013).

5.1 Contributed – Medical/Clinical Trials

Wednesday 4 September 9am - 10.20am

Quantifying the bias in estimated treatment effects if an early stopping rule is used in clinical trials: an analytic assessment

Stephen Walter
McMaster University, Hamilton, Ontario, Canada

Objectives

Stopping rules (SRs) for clinical trials control Type I error rates when interim analyses are planned, and permit early identification of the superior treatment. However, little attention has been paid to the associated bias in the estimated treatment effect. The objective of this work is to analytically quantify this bias, and other relevant quantities.

Methods

We describe trials with a continuous outcome variable, and three common SRs (Pocock, O'Brien-Fleming, or Haybittle-Peto). General expressions for the target quantities are derived and evaluated to obtain numerical results for each SR, and with one or two-sided significance testing. Results are computed for various typical scenarios.

Results

While overestimation in stopped studies can be substantial, the bias among all studies with an SR is typically less than 15% of the true treatment effect, but can be larger. Underestimation in studies that do not stop early is typically small. The probability of stopping depends strongly on the true effect size and the sample size, and can be substantial. The relative weight of stopped studies in a meta-analysis can also be large. The specific SR selected is a determinant of the importance of each of these effects.

Investigators wishing to estimate the treatment effect (in addition to simply identifying the superior treatment) should be aware of the potential bias if they adopt an SR, and that over-estimation may be substantial if the SR is invoked. Meta-analysis of studies that incorporate an SR will be less affected by bias.

5.1 Contributed – Medical/Clinical Trials

Wednesday 4 September 9am - 10.20am

Use of historical information to supplement a future study: opportunity and difficulty

Nicholas Galwey
GlaxoSmithKline, Stevenage, UK

There is currently strong interest in the possibility of utilising historical placebo (or other control) data to supplement the data obtained in a new clinical trial, thereby increasing its power and/or reducing its size. It is generally considered that historical data should be discounted relative to data from the new trial, but determination of the appropriate degree of discounting is a major difficulty. This presentation provides methods for assessing the consequences of different choices or decisions.

The discount can be expressed in terms of the difference between the historical and future means, represented by an *offset variance component* tau^2. In a Bayesian analysis, there must be a strong prior belief that tau^2 is small if much value is to be obtained from the historical data, but a range of such priors can be explored in a sensitivity analysis. Alternatively, the same numerical results can be viewed in terms of minimisation of the mean square error when biased historical data are included in the estimate, and a range of values of a bias parameter |*d*| can be explored. The choice of value for tau^2 or |*d*| can have a substantial effect on the inferences made, yet a wide range of values may be consistent with the data. Fortunately, external information can sometimes provide additional guidance on the appropriate value. On this basis, it is likely that historical data can provide a valuable contribution to exploratory studies in drug development, where bias can be accepted as one of the research project's risks.

5.2 Invited – RSS medal/prize winners & Contributed – Methods & Theory

Wednesday 4 September 9am - 10.20am

Estimation and hypothesis testing in high-dimensional transposable data

Anestis Touloumis[1,3], Simon Tavare[2,3], John C. Marioni[1]
[1]EMBL-European Bioinformatics Institute, Hinxton, UK, [2]University of Cambridge, Cambridge, UK, [3]Cancer Research UK Cambridge Institute, Cambridge, UK

Transposable data refer to random matrices where both the rows and the columns correspond to features of interest and dependencies might occur among and between the row and column variables. For example, consider a cancer study where for each subject gene expression levels are measured in multiple tumour fragments and the tumour fragments satisfy a spatial and/or temporal order. For each subject, we can write the data in a matrix form where the row variables correspond to genes and the column variables to tumour fragments. Interest might lie in drawing inference about the gene expression levels and the dependence structure between the genes and the tumour fragments.

A model for transposable data is the matrix-variate normal distribution which describes the dependence structure as the Kronecker product of two covariance matrices, one for the row and one for the column variables. In this talk, we present shrinkage estimators for these covariance matrices, we discuss their properties and we compare them via simulation to penalized maximum likelihood estimators. Further, we propose tests for the identity and sphericity hypothesis for the row covariance matrix while treating the column covariance matrix as ‘nuisance'. The two proposed tests are nonparametric and they do not specify the matrix-variate distribution of the high-dimensional transposable data. In simulations, the good performance of the proposed tests is verified and we illustrate the above using an empirical example.

5.3 Invited – RSS Medal winner & Contributed – Design of experiments

Wednesday 4 September 9am - 10.20am

Optimal designs for two-parameter nonlinear models with application to survival models

Alan Kimber, Maria Konstantinou, Stefanie Biedermann
University of Southampton, Southampton, Hampshire, UK

Censoring occurs in many industrial or biomedical 'time-to-event' experiments. Finding efficient designs for such experiments can be problematic since the statistical models involved will usually be nonlinear, making the optimal choice of design parameter dependent. We provide analytical characterisations of locally D- and c-optimal designs for a class of models, thus reducing the numerical effort for design search substantially. We illustrate our results using the natural proportional hazards parameterisation of the exponential regression model. Different censoring mechanisms are incorporated and the robustness of designs against parameter misspecification is assessed. Links between our results and the Cox proportional hazards model are considered.

5.3 Invited – RSS Medal winner & Contributed – Design of experiments

Wednesday 4 September 9am - 10.20am

Bayes linear kinematics in the design of experiments

Malcolm Farrow[1], Kevin Wilson[2]
[1]Newcastle University, Newcastle upon Tyne, UK, [2]University of Strathclyde, Glasgow, UK

We consider the choice of an experimental design by maximising the prior expectation of a utility function, within a Bayesian framework. The computation required for this maximisation can be very demanding. In particular, in problems where computation of posterior expectations would require intensive methods such as Markov chain Monte Carlo, the design calculations can require repeated simulations. This paper offers an alternative approach, applied to the particular case of the design of experiments involving groups of binomial trials, based on Bayes linear kinematics. Conjugate prior distributions are given to the binomial parameters for different groups. These are then linked through a Bayes linear Bayes structure. The resulting solution, found using Bayes linear kinematics, does not require numerically intensive or simulation-based methods and hence greatly reduces the computational burden. We also introduce a Bayes linear kinematic utility function which measures the benefit of performing an experiment in terms of information gain. The approach is illustrated with two applications, one concerned with software usability testing and the other with bioassay.

5.4 Invited – YSM 2013 Prize winners

Wednesday 4 September 9am - 10.20am

Rolling Markov Chain Monte Carlo

Din-Houn Lau, Axel Gandy
Imperial College London, London, UK

This presentation introduces a dynamic system that controls the accuracy of an estimate of a model as new data are observed. The system involves a Markov Chain Monte Carlo (MCMC) method, which is not restarted after new observations are revealed; hence the term rolling. This rolling MCMC can be paused and un-paused depending on the accuracy of the estimate. The performance of the system is demonstrated using a state space model for predicting the end of season ranks of the English Football Premier League.

5.4 Invited – YSM 2013 Prize winners

Wednesday 4 September 9am - 10.20am

X-11 versus SEATS: a comparative study of revisions on short time series

Folasade Ariyibi
Office for National Statistics, Newport, UK

When carrying out a seasonal adjustment, for greater quality of results, at least five years of data is essential to start with. On rare occasions it is necessary to perform seasonal adjustment on series which only have a span of three years. Very short series such as these are likely to have large revisions as new data can greatly change the estimates of the seasonal factors. To maximise the quality of seasonally adjusted estimates it is useful to know whether a particular method of seasonal adjustment is preferred in these circumstances. This presentation outlines an investigation testing which seasonal adjustment methods provide smaller revisions when dealing with short time series. The two seasonal adjustment methods being considered are X-11 and SEATS. The X-11 algorithm uses a non-model based approach whilst SEATS uses a model based approach.

5.4 Invited – YSM 2013 Prize winners

Wednesday 4 September 9am - 10.20am

Modelling benefits and harms of mammographic screening for breast cancer using a Bayesian cohort simulation model

Necdet Gunsoy[1], Montserrat Garcia-Closas[1], Sue Moss[2]
[1]Institute of Cancer Research, Sutton, Surrey, UK, [2]Queen Mary University of London, London, UK

Since the introduction of population-wide mammographic screening, the incidence of breast cancer has increased considerably in the UK, and has raised concerns over the potential for overdiagnosis due to screening, i.e. cancers that would never have been diagnosed in the absence of screening. Whilst it is generally accepted that screening reduces mortality from breast cancer, the balance between this benefit and the harms of overdiagnosis remains widely debated.

We developed a Bayesian cohort simulation model incorporating breast cancer progression and survival, breast cancer incidence, and data on past coverage and uptake of mammography screening and use of hormone replacement therapy. We estimated key breast cancer progression parameters through calibration to the age-specific incidence of invasive and in-situ breast cancer from 1971-2010 in the UK for a cohort born in 1935 and screened from ages 50-64 years. The progression of identical cohorts was simulated in the absence of screening and for undergoing triennial screening from age 47-73 years.

Compared to an unscreened cohort, the cumulative incidence of breast cancer was 2.4% (95% Bayesian credible interval: 1.5%-3.9%) higher in the cohort undergoing screening according to past policy, and 4.5% (2.9%-7.1%) higher in a cohort undergoing triennial screening from age 47-73 years. Cumulative breast cancer mortality was 15% (11%-19%) lower in the historical and 21% (17%-25%) lower in the current screening cohort compared to an unscreened cohort.

The use of Bayesian cohort simulation methods was effective for estimating breast cancer progression parameters and screening outcomes with only the use of population-level input parameters.

5.5 Contributed – Census & Surveys

Wednesday 4 September 9am - 10.20am

Quality and compromise: what is the use of census?

Bernard Baffour[2], Thomas King[1]
[1]Newcastle University, Newcastle upon Tyne, UK, [2]University of Queensland, Brisbane, Australia

Population census has been a standard component of demographic knowledge in developed countries for two centuries but its role has changed dramatically in that time. More information has been collected on the structure of households and the attributes of the members but data quality has also been transformed. Against standard frameworks of quality, we show that developments such as usual residence or dual system enumeration represent compromises between quality dimensions. These can be seen in the context of how the use of census has changed from disseminating population totals to providing geodemographic proportions which can be integrated with survey data and analysed externally. Specific issues of demographic change, administrative data sources, international comparison, response modality etc. continue to shape the quality profile. Where there was once no argument or alternative to census it is now essential to identify the marginal information gain for specific uses through a framework such as InfoQ. Census remains integral to statistical systems worldwide although it increasingly represents information infrastructure for users rather than a source of population statistics.

5.5 Contributed – Census & Surveys

Wednesday 4 September 9am - 10.20am

Estimating the population from aggregated administrative counts: initial findings on how models perform over time

Domenica Rasulo, Martin Ralphs
Office for National Statistics, London, UK

The system for providing population and socio-demographic statistics for the UK has been built around having a census. Whilst there is broad support from users and other stakeholders for the census, it is clear that the time is right in the UK for a fresh look at all of the possible options for producing these types of data. The Beyond 2011 Programme has been established to carry out research on the options and to recommend the best way forward to meet future user needs. Our paper shows some early work within the Beyond 2011 Programme specifically focused on modelling population estimates using aggregate administrative data such as the Patient Register, the School Census and the Higher Education Statistics Agency Student Database. We aimed to identify how well these data, along with auxiliary information such as the socio-economic characteristics of local authorities, were able to estimate age-sex population counts and to explore the data requirements and implications of these for practical implementation. We investigated linear regression models, where administrative data and other auxiliary information were fitted to Mid-Year Estimates, and multilevel techniques which adjusted additionally for within-group effects by accounting for the hierarchical structure of the data. Our results indicated that while multilevel modelling represented the optimal method to estimate age-sex population counts at the local authority level using aggregate administrative data, there were significant practical implementation issues relating to data requirements. We illustrate these issues with some examples and discuss their implications.

5.5 Contributed – Census & Surveys

Wednesday 4 September 9am - 10.20am

Rethinking households – using administrative data to count and classify households with some geographical applications

Les Mayhew[1,2], Gill Harper[1,3]
[1]Cass Business School, London, UK, [2]MHA Ltd, London, UK, [3]geocreate Ltd, London, UK

Households rather than individuals are being increasingly used for research and to target and evaluate social and economic policy in fields ranging from housing markets, social investment, utility consumption, health inequalities, and the design and delivery of services in education, health or social services. As a result accurate and timely household level statistics have become an increasing necessity across a range of uses in both the private and public sectors especially at local level. However, present sources of information on households are fragmented with significant gaps and inaccuracies that limit their usefulness. In this paper, we critique present statistical arrangements, pointing to the poor quality of much of the underpinning population data especially in areas of high population churn. We then describe a new approach to data collection and household classification based on local administrative sources. The result is a more integrated system that can not only re-create the existing official CLG household typology but also enrich it, as well as provide much greater flexibility in terms of geography. The utility and advantages are demonstrated using examples based on work for the six Olympic London Boroughs during 2011. This work includes a comprehensive enumeration of households broken down into a new household typology which is also described in the paper. The result is a data set with a much greater range of household variables that can be linked to other administrative data sources as required as is demonstrated by way of examples.

5.7 Contributed – Data Science

Wednesday 4 September 9am - 10.20am

Power calculations for N-of-1 studies

Sharon Xiaowen Lin, Peter Smith
Southampton Statistical Science Research Institute, University of Southampton, Southampton, UK

N-of-1 studies are used to pilot a new intervention on one individual. In such studies, repeated measures are taken from the individual when using and not using the intervention over a period of time. A regression model with auto-correlated errors is used to access the intervention effect size. Typical problems when analysing N-of-1 studies are the small number of repeated measures, their autocorrelation and the non-normality of the responses. As a result bootstrap tests have been advocated in the literature to tackle these issues.

Motivated by a project to deliver health intervention[1] using mobile phones, we access the power of the test for an intervention effect. In this paper, we calculate the power of a bootstrap test and a Wald test for various intervention effect sizes, length of studies and autocorrelations. Our results will guild investigators as to the length of studies required to detect an effect from a new health intervention and allow them to make an informed decision on empirical designs of the N-of-1 studies, such as for how long to run the study or whether it might be better to follow two individuals for a shorter period of time.

[1]The sample data is taken from UBhave, an EPSRC foundered project, http://ubhave.org/about.html. The funding of this research project is gratefully acknowledged. We would also like to thank Lucy Yardley, Leanne Morrison, Laura Dennison, Charlie Hargood, Scott Lloyd, Derek Johnston, Marie Johnston, Paul Roderick, Elizabeth Murray, Susan Michie, Paul Little, Mark Weal for their contributions.

5.7 Contributed – Data Science

Wednesday 4 September 9am - 10.20am

A new class of models for rating data

Marica Manisera, Paola Zuccolotto
University of Brescia, Brescia, Italy

Objectives

The aim of this contribution is to present a new class of models, called Nonlinear CUB (NLCUB) models, for modelling rating data, given by the observed responses of subjects to survey questions on their latent perceptions and evaluations.

Methods/Models

NLCUB models generalise standard CUB, which uses a mixture of two random variables to model the latent components driving the observed responses on a rating scale. Nonlinear and standard CUB are special cases of a general model explaining the cognitive mechanism that, by a step-by-step decision process, drives the individuals' responses. An interesting feature of the proposed model is given by the transition probabilities, useful to describe the individual's mental stance towards the response scale. NLCUB are able to model processes with non-constant transition probabilities: the probability of increase of one rating point in the next step of the process varies depending on the rating reached at the current step.

Results/Conclusions

Results of simulation studies investigating the functioning of the estimation method proposed for NLCUB are encouraging and show that NLCUB models are able to detect the nonlinear structure of the decision process, i.e. the non-constantness of the transition probabilities. The nice features of NLCUB models are also presented by application to illustrative and real data.

5.7 Contributed – Data Science

Wednesday 4 September 9am - 10.20am

A multiple imputation approach to remove residual confounding through coarse data models

Robert Grant
St George's, University of London, London, UK

Background

Residual confounding is a major problem in analysis of observational data, occurring when a confounding variable is measured coarsely (censored, heaped, missing, etc.) and hence cannot be fully adjusted for by the usual statistical means such as multiple regression. The analysis of coarse data has been investigated by Heitjan and Rubin but methods for coarse covariates are lacking.

Methods

I propose a flexible method combining coarse data models and fully conditional specification multiple imputation to provide unbiased estimates of the fully adjusted effects, with enlarged confidence intervals which incorporate the additional uncertainty arising from the coarsening. Conditional distributions of the true values of the confounding variable can be calculated from its relationship with other variables in the dataset and the coarsening mechanism.

Results

The method is illustrated in three scenarios with linear, logistic and Cox regression, having different forms of coarsening (interval-censoring, heaping within intervals, and multilevel heaping to a single midpoint, respectively) and using different computational techniques (Tobit-like censored regression, MLE and MCMC respectively). The true effect size is successfully recovered in each scenario.

Conclusion

Despite a widespread belief that nothing can be done to help residual confounding, it can be removed using multiple imputation, provided that one can accurately model the coarsening mechanism and specify variables correlated with the confounder. Further work is needed on practical advice for diagnosing coarsening mechanisms, semi-parametric models and triply robust estimation.

Plenary 2 – Is statistics good for your health? Minding your Ps and Rs

Wednesday 4 September 10.30am - 11.20am

Doug Altman
Centre for Statistics in Medicine, Oxford, UK

Health related research affects large numbers of people. The design, analysis and reporting of that research are inherently statistical although much medical research does not involve statisticians. Errors in research conduct or reporting are widespread; they can harm future patients in tangible ways. I will illustrate these possibilities with many case-studies from across the spectrum of medicine relating to randomised trials, systematic reviews, and observational studies of various types. I will delineate various problems beginning with P and possible remedies beginning with R.

Doug Altman is director of the Centre for Statistics in Medicine in Oxford. He has published over 600 peer reviewed articles, many aimed at clarifying statistical ideas for medical researchers. He is author of Practical Statistics for Medical Research. His varied research interests include the use and abuse of statistics in medical research, studies of prognosis, systematic reviews and meta-analysis, randomised trials, and studies of medical measurement.

Doug is senior statistics editor at the BMJ and co-editor-in-chief of Trials. He is actively involved in developing guidelines for reporting research, including CONSORT, STROBE, and PRISMA, and in 2006 founded the EQUATOR Network which seeks to improve the quality of scientific publications by promoting transparent and accurate reporting of health research.

6.1 Invited – Quantile Regression

Wednesday 4 September 11.50am - 1.10pm

From LMS to GAMLSS – 25 years of semi-parametric quantile regression

Tim Cole
UCL Institute of Child Health, London, UK

Conditional regression quantiles, which relate points on the underlying cumulative frequency distribution of a random variable to some covariate, are usually assumed to be distribution-free and estimated by optimisation. But they can equally be distribution-based and estimated by least-squares. The LMS (λ-μ-σ) method is one such method, first developed 25 years ago with a view to simplifying the construction of growth charts, where the dependent variable is anthropometry such as height or weight and the covariate is age. Since then the methodology has been formalised and extended and it now goes under the title Generalised Additive Modelling of Location, Scale and Shape (GAMLSS).

The talk will explain the genesis and development of the GAMLSS methodology, highlighting its value for growth chart construction. With the aid of examples it will show how the underlying frequency distribution can be summarised concisely in terms of one or two covariates, and it will discuss the biological form of these relationships.

6.1 Invited – Quantile Regression

Wednesday 4 September 11.50am - 1.10pm

Transformations in quantile regression

Marco Geraci
University College London, London, UK

In statistical applications, transforming data may serve a number of purposes. In parametric regression, transformations are often applied to the response variable in hopes of meeting one or more assumptions of a simplified model that are unsupported by the untransformed data. A number of transformation families have been developed to address the violation of the standard assumptions of linear mean regression, and recently some of these have been extended to quantile regression to deal with nonlinear quantile functions. In this talk, I will discuss recent advances in the application of transforms in quantile regression. In particular, I will focus on the Aranda-Ordaz families of transformations for bounded outcomes.

6.1 Invited – Quantile Regression

Wednesday 4 September 11.50am - 1.10pm

Simple thoughts on simple Quantile Regression

Chris Jones
The Open University, Milton Keynes, UK

I will review some aspects of non- and fully-parametric approaches to quantile regression in the simplest context of conditionally independent data with a single covariate.

6.2 Invited – Probabilistic Analysis of MCMC

Wednesday 4 September 11.50am - 1.10pm

Variance bounding and geometric ergodicity of Markov Chain Monte Carlo kernels for approximate Bayesian computation

Anthony Lee, Krzysztof Latuszynski
University of Warwick, Coventry, UK

Approximate Bayesian computation has emerged as a standard computational tool when dealing with the increasingly common scenario of completely intractable likelihood functions in Bayesian inference. We show that many common Markov chain Monte Carlo kernels used to facilitate inference in this setting can fail to be variance bounding, and hence geometrically ergodic, which can have consequences for the reliability of estimates in practice. We then prove that a recently introduced Markov kernel in this setting can be variance bounding and geometrically ergodic whenever its intractable Metropolis-Hastings counterpart is, under reasonably weak and manageable conditions. We indicate that the computational cost of the latter kernel is bounded whenever the prior is proper, and present indicative results on an example where spectral gaps and asymptotic variances can be computed.

6.2 Invited – Probabilistic Analysis of MCMC

Wednesday 4 September 11.50am - 1.10pm

A general metric for Riemannian Manifold Hamiltonian Monte Carlo

Michael Betancourt
University College London, London, UK

Markov Chain Monte Carlo (MCMC) is an invaluable means of inference with complicated models, and Hamiltonian Monte Carlo, in particular Riemannian Manifold Hamiltonian Monte Carlo (RMHMC), has demonstrated impressive success in many challenging problems. Current RMHMC implementations, however, rely on a Riemannian metric that limits their application to analytically-convenient models. In this talk I discuss a new metric for RMHMC without these limitations and verify its success on a distribution that emulates many hierarchical and latent models.

6.4 Invited – Evidence-based support for ecosystem management

Wednesday 4 September 11.50am - 1.10pm

RSS Panel on Statistics for Ecosystem Change and the INQUEST project

Ron Smith
Centre for Ecology and Hydrology, Edinburgh, UK

This session reports on some activities of the RSS Panel on Statistics for Ecosystem Change (PSEC). PSEC was set up in 2007 as a multi-disciplinary group to promote the statistical dimension in current environmental issues. Initially the group considered biodiversity assessment in the light of the Convention on Biological Diversity 2010 targets and then it expanded its work to consider the concept of ecosystem services in decision making. In 2012 the group was part of INQUEST, a NERC-funded Valuing Nature Network project. Group members have published two papers in ecological journals, a joint meeting of PSEC and the International Environmetrics Society resulted in a special issue of Environmetrics with four papers from PSEC members, and there is one paper published and four in preparation from the recent INQUEST activity. These talks provide an insight into the work of PSEC with one talk on biodiversity assessment and two talks from the INQUEST work, one on regulation and the other on practical application of the ecosystem services concept.

6.4 Invited – Evidence-based support for ecosystem management

Wednesday 4 September 11.50am - 1.10pm

How should regional biodiversity be monitored?

Stephen Buckland[1], Stephen Baillie[2], Jan Dick[3], David Elston[4], Anne Magurran[1], Marian Scott[5], Ron Smith[3], Paul Somerfield[6], Angelika Studeny[7], Allan Watt[3]
[1]University of St Andrews, St Andrews, UK, [2]British Trust for Ornithology, Thetford, UK, [3]Centre for Ecology and Hydrology, Edinburgh, UK, [4]Biomathematics and Statistics Scotland, Aberdeen, UK, [5]University of Glasgow, Glasgow, UK, [6]Plymouth Marine Laboratory, Plymouth, UK, [7]INRIA, Grenoble, France

We consider quantification of biodiversity in the context of targets set by the Convention on Biological Diversity. Implicit in such targets is a requirement to monitor biodiversity at a regional level. Few monitoring schemes are designed with these targets in mind. Monitored sites are typically not selected to be representative of a wider region, and measures of biodiversity are often biased by a failure to account for varying detectability among species and across time. Precision is often not adequately quantified. We explore the elements needed in a good regional monitoring scheme, and consider the outputs that such a scheme can deliver.

6.4 Invited – Evidence-based support for ecosystem management

Wednesday 4 September 11.50am - 1.10pm

Role of environmental regulation in environmental management and valuation

Marian Scott[1], Mark Everard[2], Stephen Baillie[3], Ron Smith[4], Jan Dick[4], Hamish Trench[5], Rebecca Badger[6], Mary Christie[7]
[1]University of Glasgow, Glasgow, UK, [2]Pudmillo Ltd, [3]British Trust for Ornithology, UK, [4]Centre for Ecology and Hydrology, UK, [5] Cairngorms National Park Authority*,UK [6]Scottish Environment Protection Agency UK [7]Scottish Natural Heritage, UK*

Taking a holistic view of our environment, means that increasingly ecosystem services – which include water, soil, air, biodiversity – are being assessed, in terms of quality but also in terms of their value and importance to human society. The ecosystem approach was adopted by the Convention on Biological Diversity (CBD) and defined as 'a strategy for the integrated management of land, water and living resources that promotes conservation and sustainable use in an equitable way'. This led to the Millennium Ecosystem Assessment (2005) which showed how human wellbeing was critically dependent on the delivery of ecosystem goods and services and subsequently in the UK in the National Ecosystem Assessment (2011). In this presentation, I will draw on the recent work of the INQUEST project, which looked at the role of environmental regulation in ecosystem services assessment and in the evolution of smarter environmental regulation.

6.4 Invited – Evidence-based support for ecosystem management

Wednesday 4 September 11.50am - 1.10pm

Scale and uncertainty in the evaluation of natural capital and ecosystem services: case study within Cairngorm National Park

Jan Dick[1] , Ron Smith[2], Claire McDonald[2], Alistair and Ann MacLennan[3], Jeremy Roberts[4], Desmond Dugan[4], Rebecca Badger[5], Stephen Baillie[6], Hebe Carus[4], Mary Christie[7], David Elston[8], Mark Everard[9], Mathew Hawkins[10], Hazel Kendall[11], Eleanor Mackintosh[10], Ali McKnight[112], Kristin Olsen[13], Marion Potschin[14], Marian Scott[15], Bill Slee[16], Mike Smith[17], Paul Somerfield[18], Hamish Trench[10], Gregory Valatin[18], Will Boyd Wallis[10], Roy Haines Young[15]

[1]Centre for Ecology and Hydrology, Edinburgh, UK [2]CEH, [3]Balliefurth Farm, [4]RSPB, [5]SEPA, [6]BTO, [7]SNH, [8]BIOSS, [9]Pundamilia, [10]CNPA, [11]Westcountry Rivers Trust,[12]Agroecosystems, [13]IOD PARC, [14]Nottingham University, [15]Glasgow University, [16]JHI, [17]Forest Research, [18]PML

The role of scale and uncertainty in the evaluation of natural capital and ecosystem services was explored at two study sites in the Cairngorm National Park (a commercial farm and a nature reserve). While scale is an important consideration, in practice ecosystem service and natural capital assessment takes place within specified temporal and spatial boundaries, so, as long as this is appreciated, scale is not considered a major problem for individual ecosystem services assessments. Uncertainty, on the other hand, was a more important consideration because it is frequently not expressed in ecosystem service assessments. It is important to identify (i) uncertainty in the data to estimate ecosystem services values e.g. from sampling or using proxy variables and (ii) uncertainty in the personal preferences of individuals, especially for valuing non-monetary ecosystem services which are dependent on the history, current interests and the present, past and future needs of the individual. Transparent definition of the uncertainty, temporal and spatial scales of an ecosystem service assessment is critical in order to operationalise the concept.

6.5 Invited – Beyond 2011 – future production of population and socio-demographic statistics

Wednesday 4 September 11.50am - 1.10pm

Beyond 2011 – future production of population and socio-demographic statistics – an overview

Jane Naylor, Andy Teague
Office for National Statistics, Hampshire, UK

The system for providing population and socio-demographic statistics for the UK, like many other countries, has been built around having a census. The census provides a population count at a point in time and along with surveys and administrative sources, enables a wide range of socio-demographic statistics to be produced for small geographic areas. In recent times it has become more challenging and expensive to conduct censuses and household surveys in developed countries, partly because of the complexities associated with an increasingly mobile population but also because of the reliance on the public's willingness to take part. At the same time, users want a greater range of outputs to be available and updated more frequently in order to have a better understanding of the population and how it is changing.

The 2011 Census operation was highly successful but demonstrated again that it is becoming increasingly challenging and costly to carry out. It is clear that the time is right in the UK for a fresh look at all of the possible options for producing these types of data. The Office for National Statistics is currently looking at options for the production of population and small area socio-demographic statistics for England and Wales. The Beyond 2011 Programme has been established to carry out research on the options and to recommend the best way forward to meet future user needs. This paper will provide an overview of progress to date and future plans and challenges, specifically focusing on the methodological research.

6.5 Invited - Beyond 2011 – future production of population and socio-demographic statistics

Wednesday 4 September 11.50am - 1.10pm

Population estimation from administrative data-based models

Owen Abbott, Becky Tinsley, Cal Ghee
Office for National Statistics, Fareham, UK

Beyond 2011 is considering a range of options including census, survey and administrative data solutions. Since ‘census-type' solutions are already relatively well understood most of the research is focussing on how surveys can be supplemented by better re-use of ‘administrative' data already collected from the public.

This presentation will focus on research that has been undertaken to investigate administrative data-based models as possible providers of local authority estimates of the population by age and sex. These use anonymously linked administrative data in combination with a population coverage survey. The presentation describes the framework for producing population estimates under this approach, and will cover the design of statistical population datasets, sample design issues and the estimation framework. We will present results of both a simulation exercise, and early trial estimates to assess the likely quality of estimates arising from the different options. We will discuss these potential methods in the context of the wider set of population statistics and the opportunities this may provide. The plans for further work will also be described.

6.5 Invited – Beyond 2011 – future production of population and socio-demographic statistics

Wednesday 4 September 11.50am - 1.10pm

Integrating surveys and administrative data to estimate population characteristics

Salah Merad, Neha Agarwal
Office for National Statistics, Newport, UK

A key focus of ongoing research by Beyond 2011 is the development of a methodology for the production of socio-demographic outputs (statistics about population and household characteristics) under an administrative data based approach. This paper sets out proposals for the design of an integrated system to deliver socio-demographic outputs.

An integrated system design would use administrative data wherever possible, to produce estimates, either directly or indirectly through a model-based approach. A survey will be used for estimates that cannot be provided in this way. An initial assessment of administrative data and small area estimation methods has shown that in the absence of sufficient population and topic coverage on administrative sources there is limited scope for these elements to feature as a component of the socio-demographic system in the short term. Research is being undertaken into the application of small area estimation models to supplement surveys for the production of outputs for small areas or groups. It is likely that in the short-term a large scale survey will form the basis of a system design with the scope to give a more central role to administrative data sources in the longer term as topic and population coverage by administrative sources improves.

Results from work to look at initial survey design options will be discussed along with findings from initial research to explore the scope to further improve the system design in the longer-term by making use of small area estimation and administrative data with targeted surveys.

7.1 Invited – Methods for the developing use of electronic medical record databases

Wednesday 4 September 2.30pm - 3.50pm

Dealing with missing data in electronic medical records

Irene Petersen
UCL Department of Primary Care and Population Health, London, UK

Electronic health records are increasingly used for epidemiological and health service research. However, missing data is often an issue when dealing with this type of data. Up to now various approaches have been used to overcome these issues including complete case analysis, last observation carried forward and multiple imputation. In this presentation I will first highlight the issues of missing data in longitudinal records and provide examples of the limitations of standard methods of multiple imputation. I will then discuss different ways to deal with missing data in longitudinal records and demonstrate a new user written Stata command, two-fold FCS MI algorithm, that multiple impute longitudinal records. I will illustrate how the two-fold FCS MI algorithm works in practice and maximises the use of data available, even in situations where measurements are only made on a relatively small proportion of individuals in each time period.

7.1 Invited – Methods for the developing use of electronic medical record databases

Wednesday 4 September 2.30pm - 3.50pm

Using electronic medical record databases to evaluate drug benefits and harms: are the results valid and replicable?

David Springate, David Reeves, Evangelos Kontopantelis, Darren Ashcroft, Ivan Olier
University of Manchester, Manchester, UK

Databases of electronic medical records and in particular primary care databases (PCDs) are increasingly used in research. The largest PCDs contain full data on all primary care consultations by millions of patients over two or more decades. They provide a means for investigating important healthcare questions which cannot be practically addressed in a Randomised Controlled Trial. However, concerns remain about the validity of studies based on data from PCDs. Most work around validity has attempted to confirm individual data values within a dataset. We take a different approach and instead replicate published PCD studies in a second, independent, PCD. Agreement of results then implies that the conclusions drawn are independent of the data source (though this doesn't rule out that such as confounding by indication are commonly influencing both).

We replicated two previous PCD studies using the Clinical Practice Research Datalink (CPRD). The first was a retrospective cohort study of the effect of Beta-blocker therapy on survival in cancer patients using DIN-LINK. The second was a nested case-control analysis of the effects of Statins on mortality of patients with ischaemic heart disease using QRESEARCH.

Our analyses produced several important quantitative differences compared to the original studies, altering conclusions. These could not be fully explained by either demographic differences in the patient samples or structural differences between the datasets. Our study highlights both the caution that needs to be applied when assessing the findings from analysis of just a single database and the difficulties in performing replications of existing PCD studies.

7.1 Invited – Methods for the developing use of electronic medical record databases

Wednesday 4 September 2.30pm - 3.50pm

Implementing cluster randomised trials using electronic health records

Martin Gulliford[1], Tjeerd van Staa[2,3], Alex Dregan[1], Lisa McDermott[1,4], Gerrard McCann[2], Mark Ashworth[1], Judith Charlton[1], Paul Little[4], Michael Moore[4], Lucy Yardley[4]
[1]King's College London, London, UK, [2]Clinical Practice Research Datalink (CPRD) Division, Medicines and Healthcare products Regulatory Agency, London, UK, [3]London School of Hygiene & Tropical Medicine, London, UK, [4]Division of Community Clinical Sciences, University of Southampton, Southampton, UK

Background

Electronic patient records (EPRs) have the potential to substantively improve the quality of healthcare audit and research. However, the use of EPR data in research raises several methodological and analytical challenges in comparison to standard intervention trials. This study aimed to highlight the methodological and statistical challenges of implementing clustered randomised trials in EPRs.

Methods

Data from the Clinical Practice Research Datalink (CPRD) are employed to implement two pragmatic cluster randomised trials: reducing antibiotic prescribing in primary care and secondary prevention of stroke. Randomisation is clustered at the level of general practice and the unit of analysis is the patient. The primary outcomes are the proportion of acute RTI consultations with prescribed RTI antibiotics and changes in classes of prescribed AHT drugs at 6 and 12 months follow-up. Data analysis is performed according to the ‘intention to treat principle' using marginal and random-effects models.

Results

The implementation of cluster randomised trials (CRT) in GPRD is facilitated by comprehensive data accessibility and representativeness. The implementation of cluster randomised trials in GPRD data raise several administrative and methodological challenges, including intervention implementation and acceptance by practitioners, sample size calculation and practice randomisation, data availability, outcome measures definition, and analytical methods.

Conclusions

Electronic patient records have potential for evaluation of outcomes in pragmatic trials of reducing antibiotic prescribing for RTI and stroke secondary prevention. Better and more transparent data recording in the GPRD has the potential to expand the scope of the GPRD for health care research and practice.

7.2 Invited – Modern Applied Bayesian Statistics

Wednesday 4 September 2.30pm - 3.50pm

Bayesian history matching for oil reservoirs

Michael Goldstein
Durham University, Durham, UK

Asset management for an oil reservoir typically requires the construction of a simulator for the reservoir. The simulator takes as input a description of the geology of the reservoir, and returns as output various properties of production at the wells in the field. As the geology is only partially known, a key step in the use of the simulator is termed history matching, namely finding descriptions of reservoir geology which lead to simulator output which corresponds to historical data on production of the wells. The difficulties involved in history matching arise from the dimension of the input and output spaces, the complexity of the reservoir simulator, the discrepancies between the simulator and the reservoir and the time taken for a single evaluation of the reservoir simulator for any choice of inputs. We will describe and illustrate a Bayesian approach to history matching for this problem. The method is very general, and may be applied to a wide range of problems in which complex physical systems are represented by computer simulators.

7.2 Invited – Modern Applied Bayesian Statistics

Wednesday 4 September 2.30pm - 3.50pm

Galaxy formation: Bayesian history matching for the observable universe

Ian Vernon
Durham University, Durham, UK

The question of whether there exist large quantities of Dark Matter in our Universe is one of the most important problems in modern cosmology. This project deals with a complex model of the Universe known as Galform, developed by the ICC group, at Durham University. This model simulates the creation and evolution of approximately 1 million galaxies from the beginning of the Universe until the current day, a process which is very sensitive to the presence of Dark Matter. A major problem that the cosmologists face is that Galform requires the specification of a large number of input parameters in order to run. The outputs of Galform can be compared to available observational data, and the general goal of the project is to identify which input parameter specifications will give rise to acceptable matches between model output and observed data, given the many types of uncertainty present in such a situation. As the model is slow to run, and the input space large, this is a very difficult task.

We have solved this problem using general techniques related to the Bayesian treatment of uncertainty for computer models. These techniques are centred around the use of emulators: fast stochastic approximations to the full Galform model. These emulators are used to perform an iterative strategy known as history matching, which identifies regions of the input space of interest. Visualising the results of such an analysis is a non-trivial task. The acceptable region of input space is a complex shape in high dimension. Although the emulators are fast to evaluate, they still cannot give detailed coverage of the full volume. We have therefore developed fast emulation techniques specifically targeted at producing lower dimensional visualisations of higher dimensional objects, leading to novel, dynamic 2- and 3-dimensional projections of the acceptable input region. These visualisation techniques allow full exploitation of the emulators, and provide the cosmologists with vital physical insight into the behaviour of the Galform model.

7.2 Invited – Modern Applied Bayesian Statistics

Wednesday 4 September 2.30pm - 3.50pm

Bayesian modelling of compositional heterogeneity in molecular phylogenetics

Sarah Heaps, Tom Nye, Richard Boys, T. Martin Embley
Newcastle University, Newcastle upon Tyne, UK

In phylogenetics, alignments of molecular sequence data for a group of species are used to learn about the phylogeny (i.e. evolutionary tree) which places these species as leaves and ancestors as internal nodes. Sequence evolution down the tree is modelled by making a series of simplifying assumptions about the evolutionary process. Standard models assume that a common Markov substitution process applies to every branch on the tree and at every site in the alignment. However, many studies have shown that models which relax these (often unrealistic) assumptions of homogeneity can lead to more plausible phylogenetic inferences, as well as allowing learning about the root position. We propose a non-stationary model which allows compositional heterogeneity across branches, and formulate the model in a Bayesian framework. Specifically, the root and each branch of the tree is associated with its own composition vector, with a global matrix of exchangeabilities which applies everywhere on the tree. We consider two prior distributions: one in which the composition vectors are equicorrelated and another which assumes an autoregressive evolution of the composition vector down the tree.

One of the main benefits of the proposed model is that, unlike some other related models from the literature, inference can be carried out through a simple Markov chain Monte Carlo algorithm which does not require problematic dimension-changing moves. We study the performance of the model and priors in an analysis of an alignment for which there is strong biological opinion about the tree topology and root position.

7.4 Invited – Probability, Uncertainty & Risk in the Environment

Wednesday 4 September 2.30pm - 3.50pm

The Hazard Impact Model – translating weather forecasts to societal risk

Ken Mylne
Met Office, Exeter, UK

A weather forecast is of no use unless people make decisions based on it, and in most cases those decisions are around managing risk, either business risks or risks associated with hazardous weather. A couple of years ago the Met Office, in collaboration with the civil responder community (fire service etc.), revised the National Severe Weather Warning Service to a system of risk-based warnings. Risk is defined as a combination of probability and impact and the warnings use a risk matrix to assess the overall risk. Uncertainty is an inherent part of weather forecasting, particularly as the atmosphere is a chaotic system. Forecasters use a technique called ensemble prediction to help estimate probabilities of hazardous weather, but have much less objective guidance on the likely impact of such weather to complete the impact part of the matrix. The Hazard Impact Model is being developed within the Natural Hazards Partnership, a collaboration of UK organisations, to attempt to provide more objective estimates of impact for a range of natural hazards, and allow improved risk estimation. This talk will focus on weather risk from strong winds affecting the road network, but the partnership is also addressing risks from surface water flooding and landslides.

7.4 Invited – Probability, Uncertainty & Risk in the Environment

Wednesday 4 September 2.30pm - 3.50pm

Assessing volcanic eruption intensity from near-field lithic deposits

Jonathan Rougier, Rose Burden, Jeremy Phillips
University of Bristol, Bristol, UK

For any given volcano, the intensity of historical eruptions is highly informative about present and future volcanic risk. A physically-based forward model relates intensity to the distribution of lithic particle sizes in deposits. Statistical methods can be used to invert this model, and infer the intensity from the distribution. But for practical reasons, only the largest particles in an exposed outcrop are measured. This poses an interesting design question: how many of the largest particles should be measured, to achieve an acceptable uncertainty in the inferred intensity?

7.4 Invited – Probability, Uncertainty & Risk in the Environment

Wednesday 4 September 2.30pm - 3.50pm

Probability and uncertainty in seismic hazard analysis

Ian Main[1], Sarah Touati[1], Andrew Bell[1], Mark Naylor[1], Roger Musson[2], Richard Chandler[3]
[1]University of Edinburgh, Edinburgh, UK, [2]British Geological Survey, Edinburgh, UK, [3]UCL, London, UK

Probabilistic seismic hazard analysis has recently come under fire from a number of critics who argue that many recent events have turned out to be ‘surprises', at least in terms of comparison with official probabilistic seismic hazard maps – either because they occurred in areas with no previous recorded events or even geological signs of events (as in the ongoing Canterbury earthquake sequence in New Zealand) or were much bigger than previously thought possible (as in the 2011 Tohoku mega-earthquake in Japan).

Here we describe some recent work aimed at identifying and quantifying systematic and random sources of uncertainty in earthquake recurrence statistics. In particular we examine problems of model selection, parameter estimation, structure of uncertainties, and in the convergence (or not) of models and parameters to stable central limits.

We show that a combination of primarily power-law statistics and correlations in the data (due to earthquake triggering), allied with large fluctuations associated with a Poisson point process can lead to very long times to convergence, including different models being considered optimal at different times. Record breaking events have a significant effect on model parameters and on metrics used in model selection. As a consequence ‘surprises' are to be expected as inherent to such finite size sampling, rather than an intrinsic failure of probabilistic seismic hazard as a concept. However, we do suggest some elements of current practice that could be improved, such as the effect of expert over-confidence in the ‘characteristic' earthquake model.

7.5 Invited – The RPI and the CPI – where statistical and political issues collide

Wednesday 4 September 2.30pm - 3.50pm

Derek Bird[1], Jens Mehrhoff[2], Tony Cox[3]
[1]Office for National Statistics, [2]Deutsche Bundesbank, [3]RPI CPI user group

Throughout the world Consumer Price Indices (CPIs) can be controversial. Since CPIs are frequently used to uprate, or influence the uprating of, wages, pensions, benefits, tax brackets, rents, inflation-linked bonds and business contracts, differences between popular perceptions of inflation and what the statistics show are frequently contentious. It is crucially important that statisticians do the best job possible in compiling the data. But...this is not easy. There are many problems, legitimate differences of view over coverage, scope and methodology, some complex theoretical issues and no one way everyone accepts by which the indices should be compiled.

The UK has had more than its share of controversy in recent years with the traditional Retail Prices Index (RPI) normally showing a higher inflation rate than the Consumer Price Index (CPI – which is also the EU Harmonised Index of Consumer Prices), as a result of coverage and formula differences. No real surprise then that the coalition government's decision to change the uprating of certain benefits and public pensions from RPI to CPI resulted in both a legal challenge and an e-petition which topped the 100,000 signatures necessary to trigger a parliamentary debate. A consultation in Autumn 2012 on proposals to change one of the formulae – the Carli – used in the RPI to a formula (Jevons) most expert opinion considers superior saw a public response overwhelmingly in favour of keeping the RPI unchanged for a variety of reasons – but keeping it unchanged has resulted in the UK Statistics Authority stripping it of "national statistics" status due to its use of the Carli index which was considered to have an upward bias.

Where are we now? This session will bring together the latest news on how the ONS is responding to this difficult topic, some cutting edge statistical research from Germany and an update on the "political" issues.

Derek Bird of the ONS will reprise the background to the consultation setting out the differences between the UK's Retail Prices Index (RPI) and Consumer Prices Index (CPI), how these caused the indicators to yield different growth rates and why the relationship between them changed in 2010 when the approach to collecting clothing prices was improved. He will go on to look at the process that ONS instigated to consider options for change, drawing parallels with other NSIs' experiences, what the process generated in terms of user feedback and the outcome. Finally, he will consider the prospects for consumer price indicators going forward.

Jens Mehrhoff of the Deutsche Bundesbank introduces "The CIA (Consistency In Aggregation) approach – A new economic approach to elementary indices". A single comprehensive framework, known as "generalised means", unifies the aggregate and elementary levels. With the aid of this approach, theoretical conditions under which a particular index formula at the elementary level exactly equals the desired

aggregate index are identified, independent of the axiomatic properties. This makes it possible to render index calculation more precise, driving down biases of official price indices. It is shown that the choice of the elementary indices which correspond to the desired aggregate ones can be based on the elasticity of substitution alone. Thus, a feasible framework is provided which aids the choice of the corresponding elementary index. This is also demonstrated empirically in an application using detailed expenditure data within the COICOP group of alcoholic beverages in the UK from January 2003 to December 2011.

In spirits, consumers are more willing to substitute between different types of whisky than is the case for brandy or vodka. For both red and white wines, substitution is more pronounced for the New World than for European wines. Also, the elasticity of substitution tends to be higher for 12 cans and 20 bottles of lager, respectively, than for 4-packs. In particular, the Carli index performs remarkably well at the elementary level of a Laspeyres price index, questioning the argument of its “upward bias” – in fact, this approach suggests that the Jevons index has a downward bias in these cases.

Tony Cox, the Chair of the RPI CPI User Group, will explore the reactions to the government's decisions to change the index used to uprate pensions, benefits and, later, tax brackets from the RPI to the CPI. These reactions covered the spectrum from a straightforward objection to the use of an index that would usually deliver a lower figure, to detailed arguments about the validity of the techniques used to compile the indices. These reactions will be considered in the light of more recent developments.

8.1 Invited – RSS Medal winner & Contributed – Medical

Wednesday 4 September 4.15pm - 5.15pm

Modelling reporting delays for infectious diseases using splines

Angela Noufaily[1], Paddy Farrington[1], Doyo Enki[1], Paul Garthwaite[0]
[1]The Open University, Milton Keynes, UK, [2]Health Protection Agency, London, UK

Recent outbreaks such as bird and swine flu have generated new interest in infectious disease surveillance. It is important to detect such outbreaks in time in order to take suitable control measures. The delay between the times a sample is taken from an infected person and its identification is known as the 'reporting delay'. This delay influences the speed of detection of an outbreak. In this talk, we model reporting delays to enable a better detection of outbreaks. We model the hazard of the delay distribution using a linear combination of cubic M-splines. We then put a regression model on the hazard where the covariates of interest are the time of specimen, the seasonal variation and the incidence in the current or past week. We apply this approach to reporting delays data provided by the Health Protection Agency in England and Wales. The spline-based approach provides new insights into the mechanisms producing the delays and how their distribution evolves over time. These insights will prove helpful when allowing for delays in outbreak detection systems.

8.2 Invited – RSC 2012 Prize winners

Wednesday 4 September 4.15pm - 5.15pm

Objective Bayesian survival analysis using shape mixtures of log-normal distributions

Catalina Vallejos[1,2], Mark Steel[1]
[1]*University of Warwick, Coventry, UK,* [2]*Pontificia Universidad Catolica de Chile, Santiago, Chile*

Survival models such as the Weibull or log-normal lead to inference that is not robust to the presence of outliers. They also assume that all heterogeneity between individuals can be modelled through covariates. This article considers the use of infinite mixtures of lifetime distributions as a solution for these two issues. This can be interpreted as the introduction of a random effect (frailty term) in the survival distribution. We introduce the family of Shape Mixtures of Log-Normal distributions, which covers a wide range of shapes. Bayesian inference under non-subjective priors based on the Jeffreys rule is examined and conditions for posterior propriety are established. The existence of the posterior distribution on the basis of a sample of point observations is not always guaranteed and a solution through set observations is implemented. This also accounts for censored observations. In addition, a method for outlier detection based on the mixture structure is proposed. Finally, the analysis is illustrated using a real dataset.

8.2 Invited – RSC 2012 Prize winners

Wednesday 4 September 4.15pm - 5.15pm

Applications of copula regression models in paediatric research

Eirini Koutoumanou, Mario Cortina-Borja, Angela Wade
UCL Institute of Child Health, London, UK

Copulas are multivariate distribution functions that model the joint behaviour of response variables with known univariate marginals. They can be used to model complex relationships by incorporating covariates into the parameters of all marginal distributions and the copula dependence parameter.

This talk illustrates how copulas can enhance our understanding of multivariate relationships beyond the commonly applied univariate analyses. A brief overview of the development of copulas and their growing importance in many scientific areas is followed by two examples of usage within paediatric research: i) foetal ultrasound – head and abdominal circumference measurements and ii) visual acuity – right and left eye visual acuity measurements. We use models from the Generalised Linear Additive Models for Location, Scale and Shape (GAMLSS) class to specify the marginals in the copula models. This provides flexibility to model skewness and kurtosis, and facilitates obtaining maximum likelihood estimates. The resulting copula model in each case is then extended to incorporate age as a linear predictor in the appropriate model parameters. The talk concludes with a discussion on the role that copulas can have in creating multivariate centiles and other areas for future work that is currently being undertaken at the UCL Institute of Child Health.

8.3 Contributed – Copula models

Wednesday 4 September 4.15pm - 5.15pm

Estimation of copula models with discrete margins via Bayesian Data Augmentation

Michael Smith, Mohamad Khaled
University of Melbourne, Melbourne, Victoria, Australia

Estimation of copula models with discrete margins can be difficult beyond the bivariate case. We show how this can be achieved by augmenting the likelihood with continuous latent variables, and computing inference using the resulting augmented posterior. Our method applies to all parametric copulas where the conditional copula functions can be evaluated, not just elliptical copulas as in much previous work. Moreover, the copula parameters can be estimated joint with any marginal parameters, and Bayesian selection ideas can be employed. To demonstrate the potential in higher dimensions, we estimate 16-dimensional D-vine copulas for a longitudinal model of usage of a bicycle path in the city of Melbourne, Australia. The estimates reveal an interesting serial dependence structure that can be represented in a parsimonious fashion using Bayesian selection of independence pair-copula components. Finally, we extend our results and method to the case where some margins are discrete and others continuous.

8.3 Contributed – Copula models

Wednesday 4 September 4.15pm - 5.15pm

A regression model for the Copula Graphic Estimator.

Simon Lo[2], Ralf Wilke[1]
[1]*University of York, York, UK,* [2]*Lingnan University, Hong Kong, Hong Kong*

We suggest a pragmatic extension of the non-parametric Copula-Graphic Estimator (Zheng & Klein, Biometrika, 1995) to a depending competing risks model with covariates. Our model is an attractive empirical approach for practitioners in many disciplines as it does not require knowledge of the marginal distributions. Although non-observable and only set-identifiable in most applications, classical duration models typically impose ad hoc assumptions on their functional forms. Instead of directly estimating these distributions, we suggest a plug-in regression framework which utilises an estimator for the observable cumulative incidence curves which specification can be visually inspected. Our framework is general as it accommodates parametric, semi-parametric and nonparametric models for the cumulative incidences. We perform simulations and estimate an unemployment duration model to demonstrate the advantages of our model compared to classical duration models such as the Cox proportional hazards. Our simulations suggest that there are not negligible risks involved if wrong assumptions are made about the marginal distributions as this can result even in the wrong sign of the estimated covariate effect.

If the dependence structure between competing risks is unknown the marginal distributions of the competing risks model are not identified. Rather than attempting to assume this away, our model delivers insights how relevant this identification problem is in a regression model with covariates. There are only two regression models for competing risks models which can deliver similar insights (Chen, JRSS B, 2010, and Honore & Lleras-Muney, Econometrica, 2006) but both make direct assumptions on the marginal distributions and are therefore more ad hoc than our model.

8.3 Contributed – Copula models

Wednesday 4 September 4.15pm - 5.15pm

Approximate uncertainty with pair-copula constructions for non-Gaussian Directed Acyclic Graph models using minimum information method

Alireza Daneshkhah[1], Gholam Ali Parham[2], Omid Chatrabgoun[2]
[1]Cranfield University, Cranfield, UK, [2]Shahid Chamran University, Ahvaz, Iran

Bauer et al. (2012) propose the pair-copula construction to modelling multivariate distribution represented by a Directed Acyclic Graph (DAG) for non-Gaussian distributions. Their method also permits the inclusion of conditional independence assumptions induced by a DAG. This approach is very useful for the modelling of multivariate dependencies of heavy-tailedness and tail dependence as observed in financial data. In this paper, we develop this method using the minimum information approach proposed by Bedford et al. (2013) and developed by Daneshkhah et al. (2013). The main objectives of this paper are to show that a vine structure can be used to approximate any given non-Gaussian DAG to any required degree of approximation. The standing technical assumptions we require is that the multivariate density of DAG under study is continuous and is non-zero.

Unlike the method developed by Bauer et al. (2012), we do not use any parametric distribution of the pair-copula and the marginal distributions of the variables, and the bivariate copula can be approximated using the minimum information copula based on the available data (or experts judgements).

The approximated minimum information copula is usually represented in terms of truncated series known as the bases. We examine three different bases including Ordinary Polynomial, Orthonormal and Fourier series. Finally, we apply our method to modelling the global portfolio data from the perspective of an emerging market investor located in Brazil. The results show that the multivariate distribution approximated based on the minimum information copula are fitted far better than other previously published methods.

8.4 Contributed – Industry

Wednesday 4 September 4.15pm - 5.15pm

Predictive modelling to support Quality by Design in pharmaceuticals development and manufacturing

Julian Morris[1], Zengping Chen[2]
[1]*School of Chemical Engineering & Advanced Materials Newcastle University, Newcastle, UK,* [2]*State Key Laboratory Chemo/Biosensing and Chemometrics Hunan University, Hunan, China*

Objectives

The presentation will address a number of challenges related to the use of sophisticated process analytical measurements in Quality by Design in pharmaceuticals across the development and production life cycle, including faster scale-up, pilot scale studies and six-sigma production.

Methods

In-situ online real-time sensor-based and analytical measurements are almost inevitably subjected to fluctuations/variations and depend heavily on robust analytical and process models which are resistant to fluctuations in measurement devices, measurement probes, process variables and in other external variables. In practice it is typical to observe that variations in external variables influence spectral data in a nonlinear manner which leads to the poor predictive ability and poor calibration transfer of models built on raw spectral data. This makes the task of extracting the relevant chemical and biological information, and ultimately reliable product and process understanding well beyond being routine. Advanced statistical and chemometric methodologies will be shown to significantly enhance the building of process analytical and monitoring models. Multi-scale and kernel methods for advanced process performance monitoring will also be highlighted.

Results

Industrial case studies will show how smart multivariate statistical modelling can address a range of analytical measurement challenges critical to assured product and process ‘know-how'; a case-study involving crystallisation scale up from ½ and 2L lab developments scaled up to production plant situated in another country to the where the lab developments were carried out; and finally a benchmark case study will show the potential improvements of dynamic multivariate multi-scale statistical modelling over the classical approaches.

8.4 Contributed – Industry

Wednesday 4 September 4.15pm - 5.15pm

Estimating the occurrence of debris in nuclear reactors: a Bayesian Monte Carlo approach

Paolo Mason, Richard Overton
EDF Energy, Gloucester, UK

A Monte Carlo approach is used to estimate the occurrence within eight nuclear reactors of debris originating from a specific component of the Advanced Gas-cooled Reactor fuel assembly, namely the anti-gapping unit shim, on the basis of evidence gathered in fuel channel inspections since the stations' start of life to the present day. The debris consists of a countable amount of 'petals' (25x25mm in dimensions) which are believed to break off the shim upon loading or discharging of a fuel assembly.

A model was set up of the eight reactors with respect to the generation, detection and removal of anti-gapping unit shim debris. The model was implemented in a Monte Carlo route by means of which the conditional distribution of the model parameters (given the historical record of anti-gapping unit shim debris detection in fuel channels) was sampled via the conditional acceptance of model parameter sets drawn from a prior distribution based on engineering judgement.

From the conditional distribution of the model parameters the amount of AGU shim debris presently fouling the reactors was subsequently estimated via a deterministic route (it is concluded that the AGU shim debris does not pose a threat to safe operation).

The sensitivity of this estimate to the input data was assessed.

Some tests of the general Monte Carlo Bayes methodology used in the present study will be presented.

8.5 Contributed – Environment

Wednesday 4 September 4.15pm - 5.15pm

Exploring sources of uncertainty in the cloud model MAC3

Jill Johnson, Lindsay Lee, Ken Carslaw, Zhiqiang Cui
University of Leeds, Leeds, UK

The effect of global aerosols on clouds is one of the largest uncertainties in the radiative forcing on the climate. The complex and highly computational cloud model MAC3 with bin-resolved microphysics can be used to simulate the formation of deep convective clouds given a set of microphysical and atmospheric parameters, some of which are subject to a degree of uncertainty. In this work, we aim to identify the parameters that drive uncertainty in a set of model outputs from MAC3, and in particular, we look to quantify the cloud response to aerosol in the atmosphere and determine the factors that most contribute to it.

The computationally intensive nature of the MAC3 model means that classical methods for uncertainty analysis involving direct Monte Carlo simulation are not feasible in real time. We therefore adopt a strategy of using statistical emulation in order to explore the model uncertainty. An emulator is considerably quicker to evaluate than running the model simulator itself. Once validated, an emulator can be used to explore a model output over the full input space defined by the set of uncertain model inputs, allowing for a variance-based sensitivity analysis to be performed and the leading causes of parametric uncertainty to be identified.

Here, we determine the main sources of parametric uncertainty and the response to aerosol for a set of 12 outputs from the MAC3 cloud model, including particle masses, particle concentrations and precipitation rates.

This research is funded as part of the NERC project consortium ACID-PRUF.

8.5 Contributed – Environment

Wednesday 4 September 4.15pm - 5.15pm

Investigating species interactions by modelling multivariate time series data

Hideyasu Shimadzu[1], Maria Dornelas[1], Peter Henderson[2], Anne Magurran[1]
[1]University of St Andrews, St Andrews, UK, [2]Pisces Conservation, Lymington, UK

Understanding the interactions between species within an ecological community is a key challenge in biodiversity research. Here, we focus on monthly time series records of an exceptionally well-documented estuarine fish assemblage in the Bristol Channel. Given the multi-species time series data, we have developed a model for a multivariate feedback system in which the outputs can be the inputs and vice versa. The model assumes linear interactions between species as a tractable approximation. To examine the extent to which the abundance of a given species is driven by the other species, we have analysed the model in the spectrum domain and calculated the contribution ratio of each species at each frequency.
The result suggests that our modelling approach dealing with an ecological community as a multivariate feedback system provides new insights into species interactions. We demonstrate how it enables further analysis into ecologically relevant groups of species that underpin the functioning of the system as a whole.

8.5 Contributed – Environment

Wednesday 4 September 4.15pm - 5.15pm

Using Bayesian methods for benefit transfer from choice experiments

Jacqueline Potts[1], Klaus Glenk[2], Sergio Colombo[3]
[1]Biomathematics and Statistics Scotland, Aberdeen, UK, [2]SRUC, Edinburgh, UK, [3]IFAPA, Andalusia, Spain

Choice experiments are often used for the economic valuation of non-market goods such as ecosystem services. In an approach known as benefit transfer policymakers sometimes want to infer values for one location (the ‘policy' site) from those estimated for another location (the ‘study' site). We demonstrate how Bayesian methods can be applied in order to use preference data from a study site to update preference information elicited from a relatively small sample of respondents at the policy site. This may improve the accuracy and precision of benefit estimates for the policy site but comes at the cost of conducting a small-scale study at the policy site. To illustrate this approach we conducted a simulation study using different priors for the parameters of a multinomial logit model. The priors give varying degrees of influence to the data obtained from the study site, in terms of their impact on posterior estimates for the policy site. We compare the priors using different sample sizes to assess how accuracy and precision depend on sample size and which prior specification is most promising in this approach. The example used to form the basis for the simulation study is a choice experiment examining the benefits of a reduction in soil erosion in two catchments in southern Spain.

8.6 Contributed – Population/Migration Statistics

Wednesday 4 September 4.15pm - 5.15pm

Combining internal migration data sources in England under a Bayesian framework

Rebecca Newell[1,2]
[1]Office for National Statistics, Titchfield, UK, [2]University of Southampton, Southampton, UK

Internal migration flows are often estimated from a collection of data sources, each with their own strengths and limitations in relation to availability, quality and measurement. Unlike births or deaths, the multidimensional nature of internal migration is characterised by large amounts of uncertainty, which is seldom quantified in migration literature.

A new methodology for combining internal migration flow data from the 2001 Census and NHSCR data sources is presented. Central to the research is the development of an integrated Bayesian modelling framework, capable of combining the strengths from each of the two data sources to produce harmonised estimates of internal migration flows at the regional level in England. The methodology also provides a mechanism for describing and presenting the uncertainty in the estimates. As internal migration is the primary mechanism behind population redistribution in England, this work has potential to provide both the demographic community and official statistics with a means of better understanding and measuring uncertainty in population estimates from multiple data sources.

8.6 Contributed – Population/Migration Statistics

Wednesday 4 September 4.15pm - 5.15pm

Two nations? Trends in UK residential segregation since 1991.

Paul Jones
Sheffield Hallam University, Sheffield, UK

This paper applies a range of measures of residential segregation to UK census data. The statistical measures are intended to capture the extent to which separate groups or classes of the population live apart rather than are integrated. This is an important policy question in terms of to what extent "social mixing" of rich and poor is being achieved. The two groups considered here are "richer" and "poorer" based on various definitions of the National Statistics Socio-economic classification (NSSEC). Various permutations of separation of "richer" and "poorer" are investigated based on how categories are merged. The measures (with permutations of classifications) are applied to UK census data covering the 1991, 2001 and 2011 censuses. Changes in segregation are analysed at region and local authority level. Preliminary results indicate that whilst indices of segregation are largely stable over time, there is evidence of increased segregation rather than integration in some parts of the country.

8.6 Contributed – Population/Migration Statistics

Wednesday 4 September 4.15pm - 5.15pm

A functional data analytic approach for forecasting population size: a case study of United Kingdom

Han Lin Shang
University of Southampton, Southampton, UK

Discrete time models are often used in demography for describing the evolution of an age-specific population. They are usually considered from a deterministic viewpoint, which in practice can be quite restrictive. The statistical method we propose is a model for the case where the evolution of the population is captured by means of a projection matrix. In this population projection matrix, age-specific fertility, mortality, emigration rates and immigration counts and their corresponding uncertainties are unknown, and ought to be estimated by a statistical model, such as the functional data analysis approach used here. Illustrated by the historical data of United Kingdom from 1975 to 2009, we found that the proposed method shows reasonably good in-sample forecast accuracy for the years between 2000 and 2009. Furthermore, we produce out-of-sample population forecasts of United Kingdom from 2010 to 2024, and compare the forecasts with those produced by the Office for National Statistics.

8.8 Contributed – Modelling

Wednesday 4 September 4.15pm - 5.15pm

Improving analysis of transplant survival data using multiple imputation

Laura Pankhurst[1], Robin Mitra[2], Dave Collett[1], Alan Kimber[2]
[1]NHS Blood and Transplant, Bristol, UK, [2]University of Southampton, Southampton, UK

NHS Blood and Transplant routinely collect survival data and covariate information on all patients who undergo organ transplantation in the UK. As transplantation is often successful, a high proportion of censored observations are present, which considerably reduces the effective sample size. Moreover, as with all routinely collected patient data, missing covariate values lead to difficulties in the analysis of factors associated with transplant outcomes. We aim to evaluate the suitability of multiple imputation to improve standard analyses of transplant survival data.

Multiple imputation is relatively easy to implement and is popular with practitioners. It enables complete data methods to be used for analysis, whilst also allowing uncertainty due to the missing values to be assessed. We will use data on outcomes following kidney transplantation to explore how multiple imputation improves the predictive ability of a model. By comparison with a complete case analysis, the impact on hazard ratios and the width of the associated confidence intervals for key prognostic factors will be assessed. Specifically, we explore the imputation of values of two covariates, height and weight, which are known predictors of graft survival. These covariates were not routinely collected in the kidney transplant patient data until 2003 and, even now missing values still occur, so values of these covariates are missing at random and by design.

8.8 Contributed – Modelling

Wednesday 4 September 4.15pm - 5.15pm

How do childhood diagnoses of type 1 diabetes cluster in time?

Colin Muirhead[1], Timothy Cheetham[1], Simon Court[1], Michael Begon[1,2], Richard McNally[1]
[1]Newcastle University, Newcastle upon Tyne, UK, [2]University of Liverpool, Liverpool, UK

Objectives

Previous studies have indicated that type 1 diabetes may have an infectious origin. The presence of temporal clustering – an irregular temporal distribution of cases – would provide additional evidence that occurrence may be linked with an agent that displays epidemicity. We tested for the presence and form of temporal clustering using population-based data from northeast England.

Materials and methods

The study analysed data on children aged 0-14 years diagnosed with type 1 diabetes during the period 1990-2007 and resident in a defined geographical region of northeast England (Northumberland, Newcastle upon Tyne and North Tyneside). Tests for temporal clustering by time of diagnosis were applied using a modified version of the Potthoff-Whittinghill method.

Results

The study analysed 468 cases of children diagnosed with type 1 diabetes. There was highly statistically significant evidence of temporal clustering over periods of a few months and over longer time intervals ($p<0.001$). The clustering within years did not show a consistent seasonal pattern.

Conclusions

The study adds to the growing body of literature that supports the involvement of infectious agents in the aetiology of type 1 diabetes in children. Specifically it suggests that the precipitating agent or agents involved might be an infection that occurs in "mini-epidemics".

Plenary 3 – Significance Lecture – The Bayesian Revolution

Wednesday 4 September 5.25pm - 6.15pm

Sharon Bertsch McGrayne

From spam filters and machine translations to locating Air France 447, Bayes' rule pervades modern life. It provides a systematic way to make decisions in the midst of uncertainty, and it commits us to updating our prior thoughts about a situation whenever new information appears. Alan B. Krueger, chair of President Obama's Council of Economic Advisers, volunteered, "It is important in decision making – how tightly should you hold on to your view and how much should you update your view based on the new information that's coming in. We intuitively use Bayes's rule every day."

Bayes' rule was published 250 years ago but, for most of the 20th century, it was deeply controversial, almost taboo among academics. Starting with its history, my talk will range over diverse fields and highlight some outstanding examples of Bayes in action. The talk will be based on my recent book, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines & Emerged Triumphant from Two Centuries of Controversy* (Yale University Press 2011).

9.1 Invited – RSC 2013 Prize winners

Thursday 5 September 9.20am - 10.40am

Spatially explicit capture-recapture with imperfect information on animal location

Ben Stevenson, David Borchers, Darren Kidney, Len Thomas, Tiago Marques
University of St Andrews, St Andrews, UK

Traditional capture-recapture approaches to estimating animal abundance or density ignore an obvious spatial component of capture probability; organisms close to traps are more likely to be captured than those that are far away. Spatially Explicit Capture-Recapture (SECR) methods improve upon these by accounting for the location of detected animals. One particular advantage is that they allow for animal density or abundance estimation using passive detectors (e.g. cameras or microphones) over a single sampling occasion. Distances between traps provide the spatial information required to implement SECR methods; however, in some situations, passive detectors can collect additional noisy data which further aid animal location estimation. Examples include received acoustic signal strength, precise time of acoustic signal arrival, estimated bearing to animal, and estimated distance to animal. Here, I present a new class of model that is capable of using such information to provide better estimates of animal density and abundance than what is typically possible using a standard SECR approach. I also introduce admbsecr, an R package that is currently under development and is capable of fitting such models.

9.1 Invited – RSC 2013 Prize winners

Thursday 5 September 9.20am - 10.40am

Static-Parameter Estimation in Piecewise Deterministic Processes using Particle MCMC Methods

Axel Finke, Adam M. Johansen, Dario Spano
Department of Statistics, University of Warwick, Coventry, West Midlands, UK

Piecewise Deterministic Processes (PDPs) form a class of time series models that evolve predictably in continuous time except at a countable number of random times. Discretely observed PDPs may be viewed as a generalisation of hidden Markov models. Particle MCMC methods allow for simultaneous estimation of the unknown parameters and the state of such dynamic systems by combining sequential Monte Carlo and MCMC methods. We develop non-standard particle MCMC algorithms suitable for performing inference in PDPs and provide examples.

9.1 Invited – RSC 2013 Prize winners

Thursday 5 September 9.20am - 10.40am

Clustering high dimensional data streams by feature space partitioning.

David Hofmeyr, Nicos Pavlidis, Idris Eckley
Lancaster University, Lancaster, UK

The data stream environment is garnering increasing attention due to the increased incidence of situations where traditional "offline" methods can be inappropriate. These include high frequency data sequences, non-stationary time series and big data.

Clustering is a vital task in machine and statistical learning as it allows one to build descriptive models of otherwise potentially unintelligible data. In addition, in the realm of predictive modelling, where the assumption of data homogeneity is often crucial, clustering can be used as a pre-processing method to extract homogeneous subsets from highly heterogeneous populations.

Clustering of very high dimensional data suffers from the "uniform distance" phenomenon, which renders many traditional approaches useless. Moreover, the incremental nature of "online" learning poses yet more challenges. Our work is aimed at developing a framework to combat these challenges. The approach we take is based on the principal projection divisive clustering algorithms; which create hierarchical clustering models based on a collection of simple partitioning rules in the first one dimensional eigen-subspaces. We use kernel density estimation to estimate the marginal densities in these subspaces, which we then use to find low density separators which define the boundaries of the eventual clusters. The final model is defined by a collection of convex subsets of the feature space, the truncated density within each believed to be unimodal and therefore more homogeneous.

Experimental results from extensive simulations have shown the efficacy of our approach.

9.2 Invited – Recent advances in functional data analysis

Thursday 5 September 9.20am - 10.40am

Function-valued traits in evolution

John Moriarty
University of Manchester, Manchester, UK

Many biological characteristics of evolutionary interest, such as growth curves, reaction-norms and distributions, are function-valued (or functional) data. In this context, phylogenetic comparative methods use phylogenetic trees to compare the data. I will present a model-based approach, together with a verification using synthetic data.

9.2 Invited – Recent advances in functional data analysis

Thursday 5 September 9.20am - 10.40am

Functional factor analysis for periodic remote sensing data

Chong Liu[2], Surajit Ray[1,2], Giles Hooker[3], Mark Friedl[2]
[1]University of Glasgow, Glasgow, UK, [2]Boston University, Boston, USA, [3]Cornell University, Ithaca, USA

We present a new approach to factor rotation for functional data. This rotation is achieved by rotating the functional principal components towards a pre-defined space of periodic functions designed to decompose the total variation into components that are nearly-periodic and nearly-aperiodic with a pre-defined period. We show that the factor rotation can be obtained by calculation of canonical correlations between appropriate spaces which makes the methodology computationally efficient. Moreover we demonstrate that our proposed rotations provide stable and interpretable results in the presence of highly complex covariance. This work is motivated by the goal of finding interpretable sources of variability in vegetation index obtained from remote sensing instruments and we demonstrate our methodology through an application of factor rotation of this data.

9.2 Invited – Recent advances in functional data analysis

Thursday 5 September 9.20am - 10.40am

Regression models over bidimensional manifolds

Bree Ettinger, Simona Perotto, Laura M. Sangalli
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy

We adopt a Functional Data Analysis approach and propose a regression model for data spatially distributed over general two-dimensional Riemannian manifolds. The model is a generalised additive model with a regularising term involving a suitable differential operator computed over the non-planar domain. We show that the estimation problem can be solved first by conformally parametrizing the non-planar domain and then by generalising existing models for penalised spatial regression over planar domains. The conformal parametrization and the estimation problem are computed by resorting to finite element methods. The estimators are linear in the observed data values and classical inferential tools are derived. The applications driving this study come from problems in cardiovascular research and in the neurosciences.

Funded by research program Dote Ricercatore Politecnico di Milano – Regione Lombardia, project: Functional data analysis for life sciences, and by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, FIRB Futuro in Ricerca starting grant SNAPLE: Statistical and Numerical methods for the Analysis of Problems in Life sciences and Engineering http://mox.polimi.it/users/sangalli/firbSNAPLE.html.

9.3 Contributed – Bioinformatics

Thursday 5 September 9.20am - 10.40am

Evolutionary links between multiple species

Stuart Barber, Colleen Nooney, Walter Gilks
University of Leeds, Leeds, UK

Objective

Phylogenetics is the study of how species have evolved and how they are related to each other. Cospeciation is the joint evolution of two or more lineages that are ecologically associated, the standard example being a host and its parasite. If cospeciation has occurred, evolutionary changes in one lineage are reflected by corresponding changes in the other lineage.

Permutation methods exist to detect cospeciation in closely related host-parasite systems by testing the null hypothesis that hosts and their associated parasites evolved independently. However, evolutionary relationships are often more complicated than a simple host-parasite pairing. We investigate methods to determine whether cospeciation is reflected across three (or more) associated phylogenies. The cospeciation could be between all lineages or only reflected in some of them.

Methods

In essence, perfect cospeciation should result in a set of perfectly "balanced" phylogenies. We propose a number of test statistics to capture this balance based on correlation of evolutionary distances between interacting species in each lineage. To implement permutation tests using these statistics requires randomisation schemes designed with more than two phylogenies in mind.

Results

Simulations show that a suitably chosen statistic and randomisation scheme is able to answer more subtle questions than the simple question of "is there evidence of cospeciation somewhere in this system?” By carefully considering how to randomise the trees, we are able to address deeper questions such as "do plant A and bird C show more cospeciation than can be explained by their common links with insect B?"

9.3 Contributed – Bioinformatics

Thursday 5 September 9.20am - 10.40am

Evaluation of trans-ethnic meta-analysis approaches for fine-mapping

Jennifer Asimit[1], Konstantinos Hatzikotoulas[1], Andrew Morris[2], Eleftheria Zeggini[1]
[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

Meta-analysis across ancestrally diverse populations may increase power to detect novel loci and fine-mapping resolution of causal variants. Furthermore, the inclusion of African ancestry samples may yield further improvements due to low linkage disequilibrium and high genetic heterogeneity. We carried out an extensive simulation study to investigate the fine-mapping resolution of trans-ethnic fixed-effects meta-analysis for five type 2 diabetes loci, under various settings of ancestral composition (European, Asian, African), allelic heterogeneity, and causal variant minor allele frequency. One measure of fine-mapping resolution is to rank the SNPs by their p-values and consider the probability that the causal variant is rank 1. We find that across all regions this probability is highest when the sample composition includes all three ethnic groups (e.g. for a causal with MAF 20%, the probabilities are 0.68, 0.643 and 0.804, for Europeans, Europeans/Asians, and Europeans/Asians/Africans, respectively). An alternative measure relies on constructing credible sets of SNPs that are most likely to be causal on the basis of statistical evidence (Bayes' factor). Examination of the median number of SNPs within 95% credible sets suggests that the most diverse ancestral compositions have smaller credible sets. The averages over the different regions of the median number of SNPs in each set, for a causal with MAF 10%, are 5.4, 4.2 and 2.4, respectively, for Europeans, Europeans/Asians, and Europeans/Asians/Africans. These results suggest similar performances for the European/Asian ancestry trans-ethnic meta-analyses and the European ancestry-only meta-analyses. In addition, fine-mapping resolution may be improved via trans-ethnic meta-analyses that include African ancestry samples.

9.3 Contributed – Bioinformatics

Thursday 5 September 9.20am - 10.40am

Bayesian protein structure alignment

Christopher Fallaize[1], Peter Green[2], Kanti Mardia[3], Stuart Barber[3]
[1]University of Nottingham, Nottingham, UK, [2]University of Bristol, Bristol, UK, [3]University of Leeds, Leeds, UK

Objectives

The structural alignment of proteins is a key problem in bioinformatics. For example, the function of a newly-determined protein may be predicted by assessing its similarity with proteins of known function. This has traditionally been done via sequence alignment. However, since the structure of a protein is more highly conserved than its sequence, more accurate predictions can be made by assessing structural similarity between proteins. Thus, as more and more new protein structures are determined, the problem of protein structure alignment is becoming increasingly important.

Methods

We use methods from statistical shape analysis. In particular, we use a Bayesian model for the alignment of unlabelled shapes, where the mapping between locations on different objects is unknown and must be inferred. As a prior distribution on alignments, we use a penalty function which penalises gaps in alignments, analogous to those used in sequence alignment methods. Our model allows for both 3-dimensional structure and amino acid sequence information to be incorporated.

Results

We show that our Bayesian model permits a posterior distribution which naturally quantifies the uncertainty in the alignment of proteins, and allows alternative alignments to be explored. The relative merits of these alternative alignments can be assessed via their posterior probabilities. This is in contrast to a point estimate given by deterministic methods, which is not desirable when there are multiple plausible alignments. We also show how the inclusion of amino acid information allows a posterior distribution of the evolutionary distance between proteins to be obtained.

9.4 Contributed – Data Science

Thursday 5 September 9.20am - 10.40am

Statistical inference: always valid but often fragile

[Deduction: "No inductive inference has a logical basis. All statistical inferences have a logical basis. Therefore, no statistical inference is an inductive inference." Deductions are always valid: if their premises are true, their conclusions follow. Inductions are risky: their conclusions are never guaranteed]

Ian Hunt
EDHEC Risk Institute, Nice, France

Statisticians are unreliable inductive reasoners – just like philosophers, psychologists, economists and all other soothsayers. I see no place for inductive inference in our statistical work. In this paper, I contrast induction with deduction, argue for sticking to deduction, table some case studies of statistical deduction at work, and raise objections to my argument.

Five reasons for sticking to deduction

- First, statistical inference is best understood, and taught, as a series of valid logical steps (crudely: assumptions->model->data->conditional conclusion).
- Secondly, deduction is elementary and helps get to the bottom of things. Luckily, like Holmes, we are versed in analysing fruitful and creative premises, and explaining what follows, logically speaking.
- Thirdly, we have no special "inductive rulebook". Pseudo-inductive cant, like "statistical significance", "confidence intervals", and "hypothesis rejection", can be tacked onto deductions. But these concepts are rules of thumb, at best.
- Fourthly, our clients should make their own risky inductive steps (if any). Ultimately, they are responsible for their own reasoning and decisions.
- Fifthly, deductions cry out to be challenged by constructive and critical alternative deductions. A good challenge demonstrates the fragility of the inferences involved.

I argue that hunting deductive fragility is an art, in which modern statisticians thrive. And as professionals, we have a fiducial duty to report fragility, lest our clients leap to illicit inductions (e.g. it’s our job to say, "Under your favoured model, capital punishment reduces murder rates, but under this other plausible model, capital punishment increases murder rates.").

Case studies of critical deduction in action

- D. Freedman (salt may not be so bad after all)
- A. Gelman (check your Bayesian model)
- Challenging inferences from RSS literature (court room probabilities, child abuse assessment and hospital quality control)

9.4 Contributed – Data Science

Thursday 5 September 9.20am - 10.40am

A framework understating statistical performance

Paul Askew
Chartered Society of Physiotherapy, London, UK

This paper presents a framework for understanding, managing and presenting statistical performance data. It provides a manageable, but multidimensional way of organising a range of possible analysis and assessment to provide a rounded and balanced picture for effective understanding and communication.

This is founded on a broad scope of assessment, based around three key analytical elements: trend, benchmark and target. That is the first multidimensional perspective. Each of those three key elements are then further disaggregated in a range of appropriate ways to provide a second, more disaggregate, multidimensional analytical assessment. There is then a third more tailored and specific multidimensional disaggregation to provide additional detailed statistical insight.

This draws on practical application of this framework in a range of sectors including health, policing and education, and across the public sector spectrum including central government, regulation and local service delivery, and in both strategic and operational environments.

This provides an overall framework to manage the communication and understanding of statistical performance, and focussed on public data. It provides a framework to build a balanced communication of statistical analysis and messages, and at the same time it provides the user and recipient of statistics and statistical analysis with a framework to both understand and question the scope and content of statistical communication.

9.5 Contributed – Official Statistics

Thursday 5 September 9.20am - 10.40am

Measuring the volume of international development assistance; lessons from national accounting

Matthew Powell[1,2]
[1]*Oxford Policy Management, Oxford, UK,* [2]*Office for National Statistics, Newport, UK*

International development assistance is probably more exhaustively evaluated than any other area of public expenditure. As well as the recent rise of randomised control trials or "RCTs" in areas where they can be applied, recent years have seen continuing demand for more traditional "theory" or process based evaluations as well as simple records of what happened and lessons learned during implementation. At the same time there have also been substantial attempts to define and measure the ultimate objectives of development assistance, i.e. the Millennium Development Goals and their indicators, and also, more to classify and track development expenditure in quite meticulous detail.

These efforts are driven by a strong desire on the part of policy makers to forge a link between development expenditures and progress towards the MDGs; to answer the fundamental questions, "what would have happened if these aid expenditures had not been made?" and, "what expenditures would, ceteris paribus, be sufficient to achieve the MDGs.

It has become increasingly apparent however that we are very far from being able to answer these questions or even the simpler ones of whether individual expenditure streams represent "value for money" and whether or not aid delivery is becoming more "productive". This paper will argue that these problems are closely analogous, or in some cases identical, to those encountered in measuring the volume of non-market output in national accounts and that there are many lessons to be learned from the techniques and conventions that the national accountants have developed.

9.5 Contributed – Official Statistics

Thursday 5 September 9.20am - 10.40am

Measuring gender equality: an index for Europe

Anne Laure Humbert[1], Anna Rita Manca[1], Ligia Nobrega[1], Irene Rioboo Leston[2], Jolanta Reingarte[1]
[1]European Institute for Gender Equality, Vilnius, Lithuania, [2]Rey Juan Carlos University, Madrid, Spain

Gender equality is a central feature of EU policy since its earliest days; however, measuring this multi-dimensional concept in a composite indicator is a challenging endeavour. The European Institute for Gender Equality (EIGE) undertook to develop the Gender Equality Index to support EU policy development, which the results of which were launched in Brussels in June 2013.

This paper expands on the methodological results associated with the development of this composite indicator. It outlines how the metric of analysis was developed and discusses its properties for measuring gender equality. It then presents the results of the correlation analysis and multivariate analysis that were used to reify the overall structure of the index, based on the initial conceptual considerations. It then presents the results of the robustness and sensitivity analysis associated with the choice of the final index, based on a multi-modelling approach.

The paper concludes with a discussion of how this method can be used in other contexts where disaggregation by sex is sought.

9.5 Contributed – Official Statistics

Thursday 5 September 9.20am - 10.40am

A new approach to statistical computing at the UK Office for National Statistics

Gareth James
Office for National Statistics, Newport, UK

A new Statistical Computing Branch (SCB) was established at ONS in 2012, with a view to improving various aspects of statistical computing at the office. The creation of the branch coincided with a number of other organisational changes at the office, one of which was to bring together the methodology and Information Technology (IT) areas into one new directorate: Strategy & Standards.

Although statistical computing is, of course, not new to the office, the new branch will coordinate projects and act as a central focus for such development. One of the original drivers for the establishment of the branch was to help bridge the gap between business area requirements for statistical production systems and the IT developers writing the systems. With SCB sitting within the former methodology area, it would be able to draw upon expertise to give statistical direction to projects. A second driver was to create a resource to carry out small-scale redevelopment projects, for example rewriting parts of existing statistical production processes using standardised software and more flexible approaches while maintaining standards.

This presentation will explain more of the context for the work, and discuss the novel processes put in place to make it happen. This will include organisational and operational aspects of resourcing the work, and the methods being used to ensure new development work meets required computing standards and supports future reuse of solutions. Examples of work highlighting successes will be included.

9.5 Contributed – Official Statistics

Thursday 5 September 9.20am - 10.40am

Improving small-area employment estimates produced by ONS business surveys

Matthew Greenaway
Office for National Statistics, Newport, South Wales, UK

Employment statistics at detailed industrial and geographical levels are vitally important to understanding local labour markets and formulating government policy. The primary source of these statistics is the Business Register and Employment Survey (BRES) run by the Office for National Statistics (ONS), and a key challenge is estimating for small domains where only sparse survey data are available. This presentation outlines investigative work carried out to improve the accuracy of these small-domain estimates using a variety of sample design and estimation methods.

BRES utilises a stratified sample with sample sizes allocated to strata using a Neyman allocation. Two stages of estimation are carried out – a design-based stage which utilises Generalised Regression Estimation (GREG) and a model-based stage used to produce small domain estimates. The Neyman allocation is optimal at the overall level, but not for small domains, and a power allocation and a 'multivariate' allocation have been tested instead in an effort to decrease standard errors at the domain level. Other investigations attempted to alter the model groups used in the GREG estimation and the level at which the small domain estimation was carried out.

Changes to the sample allocation and GREG estimation made little difference to the accuracy of small-area estimates, largely because we were constrained by the level at which the survey is stratified. However improvements to the accuracy of small-domain estimates can be achieved by altering the level at which the small-domain estimation is carried out.

10.1 Invited – Visualising Multivariate data

Thursday 5 September 11.10 - 12.30pm

Bringing together geovisualisation, time geography and computational ecology: using space-time density of trajectories to visualise dynamics in animal space use over time

Urska Demsar
University of St Andrews, St Andrews, Fife, Scotland, UK

Recent developments and ubiquitous use of global positioning devices have revolutionised movement ecology. Scientists are able to collect increasingly larger movement datasets at increasingly smaller spatial and temporal resolutions. These data consist of trajectories in space and time, represented as time series of measured locations for each tagged animal. Such data are analysed and visualised using methods for estimation of home range or utilisation distribution, which are often based on 2D kernel density in geographic space. These methods have been developed for much sparser and smaller datasets obtained through Very High Frequency (VHF) radio telemetry. They focus on spatial distribution of measurement locations and ignore time and sequentiality of measurements. We present an alternative geovisualisation method for spatio-temporal aggregation of trajectories of tagged animals: stacked space-time densities. The method was developed to visually portray temporal changes in animal use of space using three-dimensional kernel density around polylines to produce volumetric display in a space-time cube. We describe the algorithm for calculation of stacked densities and a case study, where we visualise trajectories of lesser black-backed gulls, collected over 30 days.

10.1 Invited – Visualising Multivariate data

Thursday 5 September 11.10 - 12.30pm

Looking at multivariate outcomes of designed experiments in Canoco

Cajo ter Braak[1], Petr Šmilauer[2]
[1]Biometris, Wageningen University, Wageningen, The Netherlands, [2]Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

Multivariate outcomes are tantamount in genomics, ecology and many other fields. Canoco (www.canoco.com) is a software package to look at such data, both unsupervised and supervised, using (constrained) dimension reduction methods. Two dimensional scatterplots, biplots, and triplots can be made much more informative by displaying classifications, series- and case-wise pie charts in such plots.

We illustrate some of these utilities using data from designed experiments. For such data, we also like to focus on (and visualise) the main effects of experimental factors and their interactions. But interactions are often difficult to grasp, even in the univariate case. The Principal Response Curves (PRC) method (van den Brink & ter Braak, 1999), related to ASCA (Smilde et al. 2011), solves this challenge by displaying a low dimensional view of a main effect plus its interactions, rescaled to a reference level (control treatment). PRC is particularly attractive for multivariate repeated measures.

Smilde AK, Timmerman ME, Hendriks MMWB, Jansen JJ, and Hoefsloot HCJ. 2012. Generic framework for high-dimensional fixed-effects ANOVA. *Briefings in Bioinformatics* 13:524-535.

van den Brink PJ, and ter Braak CJF. 1999. Principal Response Curves: Analysis of time-dependent multivariate responses of a biological community to stress. *Environmental Toxicology and Chemistry* 18: 138-148.

10.1 Invited – Visualising Multivariate data

Thursday 5 September 11.10 - 12.30pm

Harnessing the power of maps for data visualisation

James Cheshire
UCL, London, UK

The power of maps lies in their ability to visualise complex and multivariate data in a concise way. Good design is an integral part of ensuring that any map is both representative of its underlying data and engaging for its audience. This talk will discuss a range of maps produced from population data to highlight the ways in which they have been successful and how they could have been improved. The talk will also touch on how population maps can be rapidly disseminated online to raise awareness and encourage interaction with data. The final part of the talk will focus on the software tools available for producing high quality visual outputs.

10.2 Invited – Developments in Density Estimation

Thursday 5 September 11.10 - 12.30pm

A comparative study of new cross-validated bandwidth selectors for kernel density estimation

Enno Mammen[1], Maria Dolores Martinez-Miranda[2], Jens Perch Nielsen[2], Stefan Sperlich[3]
[1]University of Mannheim, Mannheim, Germany, [2]City university London, London, UK, [3]University of Geneva, Geneva, Switzerland

Recent contributions to kernel smoothing show that the performance of cross-validated bandwidth selectors improve significantly from indirectness. Indirect cross-validation first estimates the classical cross-validated bandwidth from a more rough and difficult smoothing problem than the original one and then rescales this indirect bandwidth to become a bandwidth of the original problem. The motivation for this approach comes from the observation that classical cross-validation tends to work better when the smoothing problem is difficult. In this paper we find that the performance of indirect cross-validation improves theoretically and practically when the polynomial order of the indirect kernel increases, with the Gaussian kernel as limiting kernel when the polynomial order goes to infinity. These theoretical and practical results support the often proposed choice of the Gaussian kernel as indirect kernel. However, for do-validation our study shows a discrepancy between asymptotic theory and practical performance. As for indirect cross-validation, in asymptotic theory the performance of indirect do-validation improves with increasing polynomial order of the used indirect kernel. But these theoretical improvements do not carry over to practice and the original do-validation still seems to be our preferred bandwidth selector. We also consider plug-in estimation and combinations of plug-in bandwidths and cross-validated bandwidths. These latter bandwidths do not outperform the original do-validation estimator either.

10.2 Invited – Developments in Density Estimation

Thursday 5 September 11.10 - 12.30pm

On log-concave densities and bi-log-concave distribution functions

Lutz Duembgen[1], Ralf Wilke[2], Petro Kolesnyk[1]
[1]University of Bern, Bern, Switzerland, [2]University of York, York, UK

In the first part we briefly review some results about the estimation of log-concave densities and illustrate the methods. Thereafter we introduce the class of bi-log-concave distribution functions F, i.e. non-degenerate distribution functions such that both log(F) and log(1 - F) are concave. This class of distributions is much broader, includes distributions with multimodal densities. Estimation via maximum likelihood is no longer feasible, but other paradigms work quite well.

10.3 Invited – The Cancer Research UK ICGC Prostate and Oesophageal Projects

Thursday 5 September 11.10 - 12.30pm

Resource allocation for large cancer genomics studies

Andy Lynch
University of Cambridge

Cancer Research UK are funding two cancer genome sequencing projects as part of the International Cancer Genome Consortium. The genomes of 500 Oesophageal and 250 Prostate cancers will be sequenced as part of this initiative, and the mutations present catalogued in order to infer something about the underlying disease. I will introduce the projects and outline the reasoning behind the methodology.

The sequencing effort is being conducted in batches, with the sequencing of each batch being outsourced. In the absence of prior information, each sample is allocated an equal degree of resources in the study design. There is, however, scope to improve upon this design and I will outline the opportunities here.

10.3 Invited – The Cancer Research UK ICGC Prostate and Oesophageal Projects

Thursday 5 September 11.10 - 12.30pm

Re-pairing reads – identifying and correcting image registration issues in Illumina sequencing data

Mike L. Smith
University of Cambridge

A key step in analysis of Illumina sequencing experiments is the consistent identification of clusters across multiple images of the flow cell surface. If this registration stage goes wrong inappropriate regions of the images will be read, with a resultant drop in the quality of the data produced.

I will present evidence that registration can go awry for a subset of a flow cell lane, leading to the two sets of sequences generated during a paired-end sequencing experiment being paired incorrectly. Moreover, because this phenomenon affects only a small region of the flow cell, it can easily be missed unless quality assessment metrics are plotted in conjunction with spatial information.

In addition to making alignment more challenging, there is potential for such read-pairs to negatively impact efforts to identify structural variants. This is particularly the case when employing methods that actively use paired reads where the two ends align an abnormal distance apart, or when trying to de novo assemble a sequence.

However, the affected reads can be ‘re-paired’ in order to partner them correctly. I will demonstrate both the impact the inclusion of the improperly paired reads and the effectiveness of this re-pairing strategy.

10.3 Invited – The Cancer Research UK ICGC Prostate and Oesophageal Projects

Thursday 5 September 11.10 - 12.30pm

Signatures of mutational processes in prostate and oesophageal cancers

Ludmil Alexandrov
Wellcome Trust Sanger Institute, Hinxton, UK

All cancers are caused by somatic mutations. However, understanding of the biological processes generating these mutations is limited. The catalogue of somatic mutations from a cancer genome bears the signatures of the mutational processes that have been operative since the first division of the fertilized egg. Until now there have been no theoretical models describing the signatures of mutational processes operative in cancer genomes and no systematic computational approaches for deciphering these mutational signatures. Here, by modelling mutational processes as a blind source separation problem, we introduce a mathematical framework that effectively addresses these questions. We applied this framework to prostate and oesophageal cancers and reveal the mutational processes that shape the genomes of these cancers. Certain signatures are associated with age of the patient at cancer diagnosis, known mutagenic exposures, such as tobacco smoking, or defects in DNA maintenance, but many are of cryptic origin. The results reveal the diversity of mutational processes underlying the development of cancer with potential implications for understanding of cancer etiology, prevention and therapy.

10.3 Invited– The Cancer Research UK ICGC Prostate and Oesophageal Projects

Thursday 5 September 11.10 - 12.30pm

Molecular palaeontology of multifocal and metastatic prostate cancers

David Wedge[1], Peter Van Loo[1], Dan Brewer[2], Ultan McDermott[1], David Neal[3] Steve Bova[4] and Colin Cooper[2]
[1] Wellcome Trust Sanger institute, [2] The Institute of Cancer Research, [3] Department of Oncology, University of Cambridge, [4] Institute of Biomedical Technology, University of Tampere

The ICGC includes studies of (i) 3 multifocal prostate cancers and (ii) 10 sets of metastases arising from primary prostate cancers. For each patient, multiple samples have been DNA sequenced across the whole genome. The phylogenetic relationships between these samples and the clonal and subclonal evolution of these cancers have been deduced using a variety of statistical methods. These include the elucidation of subclonal copy number changes via haplotype inference, clustering of mutations with a hierarchical Dirichlet process and automatic tree inference drawing information from copy number, point mutations and mutation phasing. Application to multiple samples yields a wealth of information including: the relationships between different regions of multifocal tumours and between primary and metastatic tumours; the identification of subclonal expansions across time and space; detailed information on the mechanism by which tumours spread either through local expansion or metastasis.

10.4 Invited – Statistics of paleoclimate and paleo-environment

Thursday 5 September 11.10 - 12.30pm

Bayesian estimation of radiocarbon calibration curves for palaeoenvironmental chronology construction

Caitlin Buck[1], Paul Blackwell[1], Tim Heaton[1], Mu Niu[1], Paula Reimer[2], & members of the IntCal Working Group[1]
[1]University of Sheffield, Sheffield, UK, [2]Queen's University Belfast, Belfast, UK

The archaeological uses of radiocarbon dating are widely appreciated, however, the fact that radiocarbon dating is vital to the establishment of time lines for many Holocene and late Pleistocene palaeoclimatic studies and palaeoenvironmental reconstructions is less widely understood. In part, this is due to the fact that for many years the calibration curves necessary to map radiocarbon to calendar ages were estimated using measurements on known age tree-rings which extend back only to about 12,500 before present.

Calibration curves are needed since the amount of radioactive carbon in the earth's atmosphere has not remained constant over time. Statistical curve estimation is used, rather than mechanistic models, because the mechanisms that drive radiocarbon variation are complex and not sufficiently well understood.

Since 2009, the types of high quality record that supply data suitable for curve estimation has diversified beyond tree-rings to include samples from lake and ocean sediments, cave deposits and corals. Consequently, a large group of scientists (the IntCal Working Group-IWG) have sought ways to manage the various sources of uncertainty associated with the new data and provide calibration curves over the entire radiocarbon dating range i.e. to 55,000 years before present.

In this talk one of the statisticians in the IWG will talk about: the various types of palaeoenvironmental data that comprise the 2013 IntCal database, the statistical work involved in estimating the curves, features of the latest versions of the curves and current and future plans for work that is likely to benefit those interested in palaeoenvironmental reconstruction.

10.4 Invited – Statistics of paleoclimate and paleo-environment

Thursday 5 September 11.10 - 12.30pm

Estimating periodicities from time-uncertain and irregularly spaced palaeoclimate data

Andrew Parnell, Piotr Tokarski
University College Dublin, Dublin, Ireland

Many climate time series suffer from temporal uncertainty and irregular spacing. This means that standard time series analysis methods are inappropriate or give misleading answers about the properties of the data. We focus here on estimating low frequency behaviour of climate time series whilst taking into account the temporal features of the data set. We explain how temporal uncertainty and irregular spacing arises in several contexts and suggests some methods for further analysis based on Bayesian and non-Bayesian techniques. We contrast theses new versions with earlier attempts in the literature.

10.4 Invited - Statistics of paleoclimate and paleo-environment

Thursday 5 September 11.10 - 12.30pm

Abrupt changes in Greenland Ice Sheet runoff and sea water temperature since 1821 recorded by coralline algae

Nicholas Kamenos[1], Trevor Hoey[1], Peter Nienow[3], Sean O'Neill[1], Thomas Claverie[5], Anthony Fallick[4], Joans Thormar[2]
[1]University of Glasgow, Glasgow, UK, [2]University of Oslo, Oslo, Norway, [3]University of Edinburgh, Edinburgh, UK, [4]Scottish Universities Envrionmental Research Centre, East Kilbride, UK, [5]University of Massachusetts, Amherst, USA

The Greenland Ice Sheet (GrIS) contains the largest store of fresh water in the northern hemisphere, equivalent to ~7.4m of eustatic sea level rise, but its impacts on current, past and future sea level, ocean circulation and European climate are poorly understood. Previous estimates of GrIS melt, from satellite observations, temperature driven melt-models and palaeo-reconstructions over <100y, show a trend of increasing melt. There are however no runoff data of comparable duration with which to validate temperature-based runoff models, or relationships between the spatial extent of melt and runoff. Further, longer runoff records that extend GrIS melt records to centennial timescales would enable recently observed trends to be put into a better historical context. We measured Mg/Ca, $\delta^{18}O$ and structural cell size in annual growth bands of red coralline algae to reconstruct near surface sea water temperature and melt/runoff from the GrIS.

Temperature: we reconstructed sub-annual resolution record of water temperature in Disko Bugt between 1821 and 2009, showing an abrupt change in temperature oscillation patterns during the 1920s which may be attributable to an alteration in the interaction between atmospheric temperature and mass loss from Jakobshavn Isbrae. GrIS.

Runoff: we produced the first reconstruction of GrIS runoff between 1939 and 2002 and observed relationships between historic runoff, relative salinity and marine summer temperature. Our reconstruction shows a trend of increasing reconstructed runoff since the mid-1980s. We suggest that since the late 1930s atmospheric temperatures have been important in forcing runoff from the western GrIS.

10.5 Invited – Measuring Child Poverty

Thursday 5 September 11.10 - 12.30pm

Panel discussion

Peter Matejic[1], Matthew Tinsley[2], Matt Barnes[3], John Veit-Wilson[4]
[1]Child Poverty Unit, London, UK, [2]Policy Exchange, London, UK, [3]NatCen, London, UK, [4]Newcastle University, Newcastle, UK

Peter Matejic from the UK Government's Child Poverty Unit will set the scene with an overview of existing measures of child poverty and an update on the consultation process.

Matthew Tinsley will discuss poverty statistics and policy – why better measures are needed and why new, wider measures that reflect the multi-dimensional and complex causes of poverty for both adults and children would be better than the current, income-based measures. This will draw on *Tackling the Causes of Poverty,* Policy Exchange's response to the government's child poverty consultation, which noted that poverty is regularly passed down generations, rather than focusing purely on early years and life chances.

Matt Barnes will contribute on *Poverty in Perspective,* giving an overview of how NatCen/Demos also adopted a multi-dimensional approach, to better understand the lived experience of poverty and to generate new insights into how to tackle it.

Professor John Veit-Wilson will pick up a number of important questions raised in responses to the consultation. For example, is the government failing to distinguish between measures of poverty, the characteristics of poor children and the associations and the consequences of poverty? How will new measures fit with established measures of poverty?

11.2 Invited – Papers from the journal of the Royal Statistical Society

Thursday 5 September 1.40pm - 3.00pm

A model based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data.

Francesco Finazzi[1], Marian Scott[2], Alessandro Fassò[1]
[1]*University of Bergamo, Bergamo, Italy,* [2]*University of Glasgow, Glasgow, UK*

This paper is devoted to the development of a statistical framework for air quality assessment at the country level and for the evaluation of the ambient population exposure and risk with respect to airborne pollutants. The framework is based on a multivariate space-time model and on aggregated indices defined at different aggregation levels in space and time. The indices are evaluated, uncertainty included, by considering both the model outputs and the information on the population spatial distribution.

The framework is applied to the analysis of air quality data for Scotland for 2009 referring to the European and the Scottish air quality legislation.

11.2 Invited – Papers from the journal of the Royal Statistical Society

Thursday 5 September 1.40pm - 3.00pm

Spatial Spline Regression models

Laura M. Sangalli[1], James O. Ramsay[2], Timothy O. Ramsay[3]
[1]MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy, [2]Department of Psychology, McGill University, Montréal, Canada, [3]Ottawa Health Research Institute, Ottawa, Canada

We describe a model for the analysis of data distributed over irregularly shaped spatial domains with complex boundaries, strong concavities and interior holes. Adopting an approach typical of functional data analysis, we propose a Spatial Spline Regression model that is computationally efficient, allows for spatially distributed covariate information and can impose various conditions over the boundaries of the domain. Accurate surface estimation is achieved by the use of piecewise linear and quadratic finite elements.

Funded by research program Dote Ricercatore Politecnico di Milano – Regione Lombardia, project: Functional data analysis for life sciences, and by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, FIRB Futuro in Ricerca starting grant SNAPLE: Statistical and Numerical methods for the Analysis of Problems in Life sciences and Engineering http://mox.polimi.it/users/sangalli/firbSNAPLE.html.

11.2 Invited – Papers from the journal of the Royal Statistical Society

Thursday 5 September 1.40pm - 3.00pm

A mixed effects model for identifying goal scoring ability of footballers

Ian McHale[1], Lukasz Szczepanski[1,2]
[1]The University of Salford, Salford, Greater Manchester, UK, [2]Smartodds, London, UK

The paper presents a model that can be used to identify the goal scoring ability of footballers. By decomposing the scoring process into the generation of shots and the conversion of shots to goals, abilities can be estimated from two mixed effects models. We compare several versions of our model as a tool for predicting the number of goals that a player will score in the following season with that of a naïve method whereby a player's goals-per-minute ratio is assumed to be constant from one season to the next. We find that our model outperforms the naïve model and that this outperformance can be attributed, in some part, to the model disaggregating a player's ability and chance that may have influenced his goal scoring statistic in the previous season.

11.3 Invited – An introduction to the Omics revolution

Thursday 5 September 1.40pm - 3.00pm

The Omics revolution and the statistical challenge it poses

Sarah Lewis
University of Bristol, Bristol, UK

The revolution in genomics started with a project to uncover every gene in the human genome, natural variation in those genes has since been interrogated for associations with common genetic disease and phenotypes to identify targets for disease treatment and prevention. However, uncovering pathways for disease does not end there. Unpicking etiological processes for disease requires much more than knowledge of which genes are involved, but also an understanding of when and how such genes are activated, which cells they operate in, which proteins, hormones and other molecules they influence and the order in which these processes occur to lead to disease.

The technology now exists to produce, cheap high through-put data on epigenetic markers (epigenomics) and RNA (transcriptomics) which indicates which genes are expressed at any given time in a given tissue. It is also possible to carry-out large scale studies of proteins (proteomics) and metabolites (metabolomics). By integrating the above data, one can start to build up a more detailed picture of what is occurring in processes which lead to disease. The challenge now is not in collecting this data but in analysing it in a coherent way to produce information which may ultimately lead to new ways to intervene at an early stage to prevent disease.

11.3 Invited – An introduction to the Omics revolution

Thursday 5 September 1.40pm - 3.00pm

Data analysis issues in epigenetic epidemiology

Caroline Relton[1,2]
[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [2]Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne, UK

The rapid increase in interest in the role of epigenetics in common complex diseases has fuelled the generation of large data sets in population based studies. Epidemiological approaches for understanding the role of both the determinants of epigenetic variation and its relationship to health and disease have many merits but also raise challenges. Foremost amongst these are: (i) the requirement for stringent quality control of data; (ii) statistical considerations when dealing with analysis of epigenetic data; and (iii) the need for innovative approaches to strengthen causal inference. The latter is a particularly important issue given the vulnerability of epigenetic patterns to confounding and reverse causation and thus spurious interpretation of association studies. These issues will be considered using illustrative examples.

11.4 Invited – Energy Statistics

Thursday 5 September 1.40pm - 3.00pm

Modelling extreme events in power system reliability analysis

Chris Dent
Durham University, Durham, UK

This presentation will describe the different classes of extreme event in power system analysis, and methods which I and others have used to model them. In particular, generation adequacy (the risk of absolute shortages of generating capacity), system flexibility (whether systems have sucient flexible resources to match the variations in demand and wind output), and system security (resilience of systems to sudden fault events) will be discussed, along with associated issues of limited data, definition of risk indices, and the relationship between models which are used and the real systems which they attempt to represent.

11.4 Invited – Energy Statistics

Thursday 5 September 1.40pm - 3.00pm

Forecasting locally stationary series with energy applications

R Killick[2], M Knight[1], I Eckley[2], G Nason[1]
[1]University of Bristol, Bristol, UK, [2]Lancaster University, Lancaster, UK

The partial autocorrelation function is an essential tool of time series analysis often used for autoregressive order determination. We introduce a new estimator for the local partial autocorrelation function for locally stationary time series and show how this estimator can be used as part of a forecasting strategy for these kinds of series. We will show how this new estimator and forecasting strategy can be used in energy applications.

11.4 Invited – Energy Statistics

Thursday 5 September 1.40pm - 3.00pm

Multidimensional covariate effects in spatial and joint extremes

Philip Jonathan[1], Kevin Ewans[2], David Randell[1], Yanyun Wu[1]
[1]Shell Projects and Technology, Chester, UK, [2]Shell Projects and Technology, Kuala Lumpur, Malaysia

Careful modelling of covariate effects is critical to reliable specification of marine and coastal design criteria. We present a spline based methodology to incorporate spatial, directional, temporal and other covariate effects in extreme value models for environmental variables such as storm severity. For storm peak significant wave height events, the approach uses quantile regression to estimate a suitable extremal threshold, a Poisson process model for the rate of occurrence of threshold exceedances, and a generalised Pareto model for size of threshold exceedances. Multidimensional covariate effects are incorporated at each stage using penalised tensor products of B-splines to give smooth model parameter variation as a function of multiple covariates. Optimal smoothing penalties are selected using cross-validation, and model uncertainty is quantified using a bootstrap resampling procedure. The method is applied to estimate return values for large spatial neighbourhoods of locations, incorporating spatial and directional effects. Illustrations of extensions to incorporation of extremal spatial dependence, and application of the spline methodology to multivariate extremes using the conditional extremes model are given.

Plenary 4 – Changing face of statistics

Thursday 3 September 3.20pm - 4.20pm

Sheila M. Bird
MRC Biostatistics Unit, Cambridge, UK

After a lexicon of changes in statistical science, public and parliamentary understanding of statistics, and in how statisticians are educated and interact with other disciplines, whither statistics? In the centenary year of the Medical Research Council, and as MRC Biostatistics Unit embarks on its own centenary celebrations, my reprise and vision are cussedly, and gratefully, those of a biostatistician.

POSTERS

Poster viewing A

Tuesday 3 September 1.30pm - 2.30pm

UK Biobank as a resource for identification of spatial variation in hearing impairment and its treatment

Mark Edmondson-Jones[1], Abby McCormack[1], Heather Fortnum[1], David R Moore[2]
[1]NIHR Nottingham Hearing Biomedical Research Unit, Nottingham, UK, [2]Cincinnati Children's Hospital Medical Center, Cincinnati, USA

Objectives

In the 1980s hearing impairment was estimated to affect approximately 16% of the adult UK population increasing to over 50% by around age 70. Various factors influence hearing sensitivity, including environmental exposures. Although hearing impairment is relatively common use of hearing aids is underutilised (c. 2-3% of the UK adult population). We set out to identify characteristics of areas with relatively high prevalence of hearing impairment and those with relatively low uptake of hearing aids. Associations with ecological factors are also investigated.

Method/Models

UK Biobank (http://www.ukbiobank.ac.uk/) is a large study in which over 500,000 adults aged between 40 and 70 completed a battery of baseline tests and questionnaires between 2006 and 2010. In addition to the taking of biological samples and physiological measures, various tests were performed, including a speech-in-noise hearing test. Invitations were issued to residents within a catchment area centred on each of 22 assessment centres throughout the UK, and so coverage of the UK was incomplete. In addition to presenting raw, indirectly age-standardised prevalence rates of hearing impairment at local authority district level, we analysed these data using a spatial Bayesian ecological regression model incorporating deprivation and area classification.

Results and Conclusions

Estimated ecological factor effects illustrate an increase in prevalence with deprivation and industrialised areas have amongst the highest prevalence of hearing impairment with London amongst the lowest. Overall, hearing aid use was around 3% within this study, with only 10% of those reporting difficulty with their hearing using a hearing aid.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Hypertension and the risk of cardiovascular disease mortality in an elderly male population: a competing risks approach

Katerina Sirigonaki[3,1], Anthony Kafatos[1], Daan Kromhout[2], Manolis Linardakis[1], Antonis Bertsias[1], Joanna Moschandreas[1]
[1]University of Crete, Crete, Greece, [2]Wageningen University, Wageningen, The Netherlands, [3]Athens University of Economics & Business, Athens, Greece

The elderly constitute a population known to be susceptible to competing risks. The aim of the present study is to investigate the effect of hypertension on the risk of CVD mortality in elderly men. Participants were 253 men aged 71+ living in rural Crete in 1991 (73% with hypertension), and followed up until 2010. Using data augmentation, Cause-Specific Hazard (CSH) functions were estimated using a stratified Cox model. An interaction term was included to test for a common effect of hypertension on both CVD and non-CVD mortality. The Cumulative Incidence Function (CIF) was calculated using the naive (1 minus Kaplan-Meier) and the proper non-parametric estimator. The Fine & Grey model was used to assess the effect of hypertension on the CIF, adjusting for other CVD risk factors. There were 111 (44%) CVD deaths and 121 (48%) non-CVD deaths. Median survival was 8 years. Five and 10- and 15-year risks of CVD mortality were 17%, 32% and 40% respectively using the CIF (1-KM estimates 18%, 39%, 56%). Based on the stratified model, the hazard of CVD mortality in men with hypertension was 1.8 (95% CI 1.1 to 3.1) times that for those without hypertension. The interaction term was significant; the corresponding hazard for non-CVD mortality was 0.8 (95% CI 0.6 to 1.3). The cumulative incidence of CVD mortality was higher for hypertensive men (subdistribution hazard 1.8, 95% CI 1.1 to 3.1). Using both CSH and CIF analyses, hypertensive status appears to remain a risk factor for CVD mortality in these elderly men.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Modelling population psychometric function in a speech-in-noise task: using UK Biobank to explore links between cognition and listening

Mark Edmondson-Jones[1], Abby McCormack[1], Heather Fortnum[1], David R Moore[2]
[1]NIHR Nottingham Hearing Biomedical Research Uni, Nottingham, UK, [2]Cincinnati Children's Hospital Medical Center, Cincinnati, USA

Objectives

26% of adults aged 40 report difficulties following a conversation with background noise; this increases to 48% by the age of 70. We aim to investigate the characteristics of this decline in hearing function and in particular its association with cognitive decline.

Method/Models

UK Biobank (http://www.ukbiobank.ac.uk/) is a large cohort study which performed a detailed baseline assessment on over 500,000 UK resident adults (aged 40 to 70) between 2006 and 2010. A large subgroup performed a Speech-in-Noise (SiN) hearing test and five cognitive tests. The SiN test involved listening to spoken digit triplets played against a noisy background and keying them via a touchscreen. The spoken digits' volume was varied adaptively according to the participants' performance. Fifteen triplets were presented to each ear.

Commonly a speech reception threshold is estimated, for UK Biobank as the mean Signal-to-Noise Ratio (SNR) of the last eight triplets corresponding to an estimate of the SNR at which 50% correct responses are expected. Instead we estimated the underlying psychometric function (the probability of a correct response as a function of SNR) allowing for stimulus-independent lapses of attention. This was implemented using a Bayesian generalised linear mixed model framework, allowing for population random effects, modified for the occurrence of unobserved lapses.

Results and Conclusions

The model results describe how aspects of attention and underlying performance vary with socio-demographic and cognitive variables. Population variability is also described. Finally, the results of this modelling are used to describe the characteristics of the UK Biobank SiN test.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Bayesian inference for probability of chronic non-communicable diseases in Nagaland: a remote hill state of India

Labananda Choudhury[1], Prasanta Barman[2]
[1]Gauhati University, Guwahati, Assam, India, [2]Kohima Science College, Jotsoma, Kohima, Nagaland, India

Background

Chronic non-communicable diseases are the leading causes of both disability and death worldwide. Deaths due to communicable diseases have drastically gone down during last few decades, whereas deaths from chronic non-communicable diseases like cardiovascular diseases, diabetes and cancer are alarmingly increasing. This fact is true in Nagaland too. While people are aware about the threat of these diseases, specific risk related information can help to focus on the gravity of the threat.

Objectives

This study aims to estimate the probability of an individual being afflicted by selected chronic non-communicable disease in different segment of population of Nagaland using Bayesian methodology.

Materials and Methods

The data used here is a primary one collected through a household survey. At first, the risk factors associated with prevalence of diseases under study are identified by logistic regression technique, which quantifies the change in odds. The estimates of the probability called Bayes estimates are obtained for different mutually exclusive segment of population by considering uniform objective prior distribution. The dependence on prior distributions has been checked by a sensitivity analysis under different reasonable choices of prior.

Results

It is observed that cardiovascular disease and diabetes are the most prevalent diseases in Nagaland. Consequently the probabilities of an individual being afflicted with these two diseases are estimated separately. The analysis indicates that in most of the cases a Bayesian approach works well when compared to the classical method in regards to probability estimations.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Timothy O Olatayo, Abass I Taiwo
Olabisi Onabanjo University,Department of Mathematical Sciences, Ago-Iwoye,Ogun State, Nigeria

Forecasting is very important in all spheres of human endeavours, since predictions of future events are extremely important in decision making processes. With respect to rainfall, which is on the rise due to global climate changes, a better forecast would help governments, corporate organisations and individuals make operational, tactical and strategic decision to combat and reduce the devastating effect of flooding caused by excessive rainfall.

Generally, forecasting methods can be divided into classical – Box-Jenkins and Time Series Regression methods – and modern – Fuzzy Time Series and Neural Networking methods – and were compared using rainfall data.

The empirical results show that the modern methods gives more accurate forecasts compared to the classical methods.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Covariate shift adaptation from a viewpoint of missing data problems

Masayuki Henmi
The Institute of Statistical Mathematics, Tokyo, Japan

In a problem of supervised learning, it is usually assumed that the probability distribution of input variables for training data is the same as that for test data. However, it is not necessarily true in practice, and this situation is often called covariate shift. The commonly used method for adjustment under covariate shift is to use the density ratio of the two distributions of input variables as a weight for a loss function, and this is a typical application of density ratios. On the other hand, it has been indicated that the problem of covariate shift can be viewed as a missing data problem under MAR (Missing At Random), and discussed that the weighting by the density ratio corresponds to the inverse probability weighting by the propensity score, which is a popular method in missing data problems or causal inference. The aim of this work is to make a further investigation of this relationship by referring to the literature on missing data problems, which has been mainly developed in biostatistics and econometrics, and to consider its practical implications in the problem of covariate shift. The topics to be discussed include semi-parametric efficiency of estimation, role of the methods for handling missing data under NMAR (Not Missing At Random) in the covariate shift and so on. It is shown that this investigation helps us to develop further methods for covariate shift problems as well as to obtain a better understanding of the relationship between the two problems in machine learning and statistics.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Gaussian tree graphical models for discrimination

Guillermina Eslava-Gomez, Gonzalo Perez-de-la-Cruz
National University of Mexico, Mexico, D.F., Mexico

We consider discriminant analysis for two populations, where a graphical Gaussian model with tree dependence graph is considered on each population. We restrict to models where dependence graphs have the same tree structure, though not necessarily with the same concentration matrix. By considering a tree graph, the maximum likelihood estimator (MLE) for each concentration matrix can be expressed analytically, and by considering the same tree structure for the two concentration matrices together with the J-divergence, the estimation of the unknown structure is solvable by finding the minimum weight spanning tree (MWST).

We use the J-divergence as a measure of discrimination between two populations, and the one to be efficiently optimized by finding the MWST. By using the MLE for each concentration matrix and the MWST we get the estimated discriminant function.

We illustrate the empirical performance of the proposed and of other existing methods using some data. This example shows similar performance for the methods using tree graphs on the models, and a better one respect to quadratic discriminant analysis for small sample sizes.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Comparison of prediction capabilities of partially replicated central composite designs in cuboidal region

Polycarp Chigbu, Eugene Ukaegbu
Univ. of Nigeria, Nsukka, Nigeria

Three variations of partially replicated central composite designs have been compared. Their prediction capabilities with respect to their prediction variances are assessed using the variance dispersion graphs in cuboidal region when the axial distance is a = 1. The results show that for n_0 = 1, 2 and 3 centre points, the replicated star designs have better prediction potentials than the replicated cube designs when interest is in predicting responses in the face-centred cube. The face-centred cube is the second-order central composite design with axial distance, a = 1.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

A new Family of Gamma Distributions

Kishore Das
Gauhati University, Guwahati, Assam, India

A new Family of Gamma Distributions (FGD) has been defined and also been derived. The motivation of defining this distribution is to find a new class of gamma distributions from which generate various known, viz., chi, chi-square, exponential, Rayleigh, Maxwell-Boltzmann, half normal, Weibull, Frechet type extreme value distribution, etc., as well as many unknown distributions. Then, a few moment properties of these distributions have been studied. A family of inverted gamma distribution has also been derived. Finally, mixture of Poisson distributions has been obtained by ascribing the FGD and chi-square as mixing distribution to the parameter of the Poisson distribution.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Cohort analysis to study wastage of manpower of the employees in Dhubri district of Assam, India

Kishore Das[1], Bhanita Das[2]
[1]Gauhati University, Guwahati, Assam, India, [2]North Eastern Hill University, Shillong, Meghalaya, India

The most natural way to investigate the pattern of wastage is to observe a homogeneous group of entrants at about the same time (cohort) and note how long each remains in the organization before leaving. Length of service in our case corresponds to the life of the employees in an organization. In this paper, to study wastage pattern estimate of the probability function that an entrant leaves during an interval, survivor function, conditional probability function, hazard function and their standard errors have been studied for length of service data of educational institutions in Dhubri District using cohort. Much of the life testing theory is based on exponential distribution. From the survivor function it is clear that probability density function follows exponential distribution and in general length of service data are incomplete. Therefore, exponential distribution and truncated exponential distribution also have been considered to illustrate the applicability of length of service data. Finally, propensity to leave of an employee in an organization life table has been constructed using cohort data.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

The UK statistical policy landscape infographic

Paul Askew
Chartered Society of Physiotherapy, London, UK

This infographic presents and inter-relates the key components of the UK's statistical policy landscape. The scope includes key bodies, legislation, policy, guidance, and key contextual factors and influences. This provides a single strategic visual overview and reference source, illustrating the evolution, relationship and synthesis of those components over the last decade. This also helps to illustrate to a lay audience the extent and development of the background rigour and governance in public data management, use and communication.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

A greedy algorithm for representative sampling: rep-sample in Stata

Evan Kontopantelis
University of Manchester, Manchester, UK

Quantitative empirical analyses of a population of interest usually aim to estimate the causal effect of one or more independent variables on a dependent variable. However, only in rare instances is the whole population available for analysis. Researchers tend to estimate causal effects on a selected sample and generalize their conclusions to the whole population. The validity of this approach rests on the assumption that the sample is representative of the population on certain key characteristics. A study using a non-representative sample is lacking in external validity by failing to minimise population choice bias.

When the sample is large and non-response bias is not an issue, a random selection process is adequate to ensure external validity. If that is not the case, however, researchers could follow a more deterministic approach to ensure representativeness on the selected characteristics, provided these are known, or can be estimated, in the parent population. Although such approaches exist for matched sampling designs, research on representative sampling and the similarity between the sample and the parent population seems to be lacking. We attempt to address the issue by proposing a greedy algorithm for obtaining a representative sample and quantifying representativeness, in Stata.

Poster viewing A

Tuesday 3 September 1.30pm - 2.30pm

A new model for estimating district life expectancy at birth in India with special reference to Assam (2001, 2010)

Labananda Choudhury[1], Rajan Sarma[2]
[1]Gauhati University, Guwahati, Assam, India, [2]Darrang College, Tezpur, Assam, India

Life expectancy at birth (e_0) is considered as an important indicator of the mortality level and quality of life in a population. In India, direct estimation of e_0 through construction of life tables is not possible due to incomplete death registration. The Sample Registration System of India provides information on e_0 only for 16 states. The estimates of e_0 for the districts are not available. Most of the indirect methods for estimating e_0 are inapplicable due to the unavailability of the required data and/or failure to satisfy the associated assumptions at the sub state level. In India, the crossover of highest life expectancy from age five to age one took place during 1980-85 and this shifts the burden of the effects on e_0 from childhood mortality to infant mortality. Thus, changes in mortality in the first year of life strongly affect life expectancy at birth. This paper seeks to establish a relationship between infant mortality, and life expectancy at birth. Using this relationship, an attempt has been made to estimate e_0 for some selected districts covered by the Annual Health Survey of India, 2010-11.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Modification of brass method and indirect estimation of infant and child mortality at district level of India

Rajan Sarma[1], Labananda Choudhury[2]
[1]Darrang College, Tezpur, Assam, India, [2]Gauhati University, Guwahati, Assam, India

Background

In India, direct estimation of child mortality rates is not possible due to insufficient death registration. Brass method is the popularly used indirect technique to estimate child mortality in developing countries. Indirect estimates of the probability of dying from birth to age x, obtained by Brass method based on child survival data are sometimes unrealistic, being not monotonically increasing for increasing x.

Method

State and district level child mortality estimates obtained by Brass method from 1991 and 2001 census data were improved by logit smoothing. Using two of the smoothed child mortality estimates, infant mortality estimate is obtained by fitting a two parameter Weibull survival function.

Results and conclusion

The reliability of the estimates have been established by cross checking with the 1991 and 2001 census data on the increase/decrease of percentage of children dead. Udupi and West Kameng district has the lowest and highest infant mortality rate respectively.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

The variance: a new equivalent formula. The double variance of the normal distribution

Muawia Abdelsalam
University of Gezira, Medani, Gezira State, Sudan

The variance: a new formula

There is a question which normal distribution should be called the "standard" one. Various authors answer it differently. Starting from the works of Gauss the standard normal was considered to be the one with variance ½. Stigler (1982) goes even further and insists the standard normal to be with the variance ½π.

This paper attempts to present a new formula, a new marginal contribution in statistics; it is related to the nature of the function of the normal distribution and other probability distribution functions. The paper will answer the question of why, starting from the works of Gauss, the standard normal distribution was considered to be the one with variance ½. The new formula is an attempt to insist that Gauss' consideration can really be proved in statistics. To support this statistical claim, this paper presents a new type of variance.

The new formula is an attempt to insist that Stigler (1982) also has a strong assertion when he goes even further and insists the standard normal to be the one with the variance ½π. As is well known, Stigler's formulation for the standard normal distribution function is advantageous because it is a much simpler and easier-to-remember formula, has unit height at zero, and there are simple approximate formulas for the quantiles of the distribution.

Poster viewing A
Tuesday 3 September 1.30pm - 2.30pm

Simultaneous confidence bands for a percentile line in linear regression with application to drug stability studies

Yang Han[1], Wei Liu[1], Frank Bretz[2], Fang Wan[1]
[1]S3RI and School of Mathematics, University of Southampton, Southampton, UK, [2]Novartis Pharma AG, Basel, Switzerland

Objectives

Simultaneous confidence bands have been used to quantify unknown functions in various statistical problems. A common statistical problem is to make inference about a percentile line in linear regression. Construction of simultaneous confidence bands for a percentile line has been considered by several authors. But only conservative symmetric bands, which use critical constants over the whole covariate range (-∞, ∞), are available in the literature. The main purpose of this research is to construct simultaneous confidence bands for a percentile line over a given covariate interval which can be finite or infinite, and to compare the bands under the average band width criterion.

Methods

Methods given in this paper allow the construction of exact symmetric simultaneous confidence bands for a percentile line over a finite interval of the covariate x. Furthermore, we propose a method of constructing an asymmetric simultaneous confidence band corresponding to each given symmetric band. We illustrate the proposed methods with a real example on drug stability study.

Results and Conclusions

Comparison under the average band width criterion shows that the exact symmetric bands can be substantially narrower than the corresponding conservative symmetric bands. Furthermore, we find that asymmetric confidence bands are uniformly, and can be very substantially, narrower than the corresponding exact symmetric bands. Therefore, asymmetric bands should always be used under the average band width criterion.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Simultaneous confidence bands for a percentile line in linear regression with application to drug stability studies

Andreas Anastasiou
University of Oxford, Oxford, UK

In statistics, maximum likelihood estimation is a widely used method for estimating the parameters found in a statistical model. One of the most important properties of the Maximum Likelihood Estimator (MLE) is related to its asymptotic distribution. Under regularity conditions, this distribution is found to be normal with mean the unknown parameter θ_0 and variance $1/I(\theta_0)$, where $I(\theta_0)$ is the expected Fisher information number of θ_0.

We examine using Stein's method how close to the normal distribution, the distribution of the MLE is. Stein's method provides explicit bounds for distributional distances even under the presence of dependence. In this poster, special attention is given to the exponential family of distributions.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Comparing sources of Welsh language statistics

Sophie Larcombe
Welsh Government, Cardiff, UK

The decennial census is currently the only source of sub-local authority level estimates of the Welsh language ability of the population of Wales. This information is available down to the output area level and is used for the development of Welsh language policy by the Welsh Government.

The Beyond 2011 Programme at the Office for National Statistics (ONS) is currently considering alternatives to a traditional census in England and Wales. This project will provide an input to the Welsh Government's considerations on Beyond 2011 by helping to understand the alternative options for Welsh language data.

This poster will present the results of comparisons between the number of Welsh speakers from the 2011 Census, the Annual Population Survey and the National Survey for Wales. Comparisons are made for the total population as well as at the local authority level and for broad age groups.

National Surveys overestimate the total number of Welsh speakers in Wales compared to the Census estimate by up to 48 per cent. These differences challenge the ability of large scale national surveys to replace a traditional census as the source of small area Welsh language statistics. Potential reasons for these differences are presented and delegates are welcome to contribute their own ideas based on the analysis.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Generalising smooth transition autoregressions

Emilio Zanetti Chini
University of Rome "Tor Vergata", Rome, Italy

This paper holds with the issue of testing, modelling and estimating the (observed) dynamic asymmetry for economic and financial data. Dynamic asymmetry is defined when the average duration of the phases is the same, but the cycle undergoes a steep fall and a steep recovery until it peaks at a slower rate, while finally it starts to fall at low, accelerating rate.

I introduce a theoretically coherent variant of the Smooth Transition Autoregression (STAR). The proposed model is able to parametrize the asymmetry in the tails of the transition equation by using a particular generalization of the logistic function. The null hypothesis of symmetric adjustment toward a new regime is tested by building two different LM-type tests. The first one maintains the original parametrization, while the second one is based on a third-order expanded auxiliary regression. Three diagnostic tests for no error autocorrelation, no additive asymmetry and parameter constancy are discussed. The empirical size and power of the new symmetry as well as diagnostic tests are investigated by an extensive Monte Carlo experiment.

An empirical application of the so-generalised STAR (GSTAR) model to four economic time series reveals that the dynamic asymmetry in the transition between two regimes is a feature to take in account for economic analysis.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Do improvements to car safety save lives?

Caroline Reeves
TRL, Wokingham, Berkshire, UK

Fatalities on British roads fell noticeably from 2007 to 2010. The aim of this research was to investigate the causes of this reduction, including how improvements to vehicle secondary safety had contributed.

‘Secondary safety' refers to the protection offered by a vehicle in the event of an accident. This differs from ‘primary safety', which refers to systems which help to prevent accidents from occurring. This analysis investigated the improvement in secondary safety for both car drivers and pedestrian casualties, to determine if the change in fatal trend seen in 2007 could be attributed to changes in the car fleet. A Generalised Linear Model was used to analyse car driver and pedestrian casualties by vehicle registration year in order to identify how car secondary safety has improved during recent decades.

The results indicate that car secondary safety for drivers has improved on both low and high speed roads. This improvement has been faster since cars registered after 1990-1991 were introduced. The trend in secondary safety for pedestrians struck by cars differs to that of driver casualties: improvements to secondary safety for pedestrians appear to have decelerated in the last decade. Both the car driver and pedestrian models were unable to detect a change in trend around 2007. This indicates that either development in new cars is not making the fleet any safer than expected, or that the inflow of new cars into the fleet has not been sufficient to have had a major effect on the fatality numbers.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Rank estimation in regression

Yuankun Wang, Kanchan Mukherjee
Lancaster University, Lancaster, UK

In this talk we introduce Rank estimators and look at why it is an interesting and better alternative to the mainstream least squares estimator in the linear regression model. Computation of R-estimators is a long-standing issue in this area for which R-estimation has not been popular in the robust estimation literature despite its attractive theoretical properties. We propose an algorithm to compute Rank estimators. The algorithm is then applied to several real as well as simulated datasets for both simple and multiple linear regression models and the results are discussed. Finally, further plans for the extension of my work are discussed.

Poster viewing B

Wednesday 4 September 1.30pm - 2.30pm

Efficiency of Neyman allocation procedure over equal and proportional allocation procedures in stratified random sampling

Olaniyi Mathew Olayiwola[1], S.A. Mustapha[2], Oladiran Hammed Bisira[3], F.S. Apantaku[4]

[1]Department Of Statistics, College Of Natural Sciences, Federal University Of Agriculture, Abeokuta, Nigeria, [2]C.R.C, Oyo, Nigeria, [3]Lagos State Polytechnic, Department Of Mathematics, Lagos, Nigeria, [4]Department Of Statistics, College Of Natural Sciences, Federal University Of Agriculture, Abeokuta, Nigeria

In sampling, we have interest in precision and we make use of prior knowledge of the population. We try to put the population into series of homogeneous groups to increase the precision. When the population of interest can be divided into *k* homogeneous groups and the sample of observation is taken from each group (stratum), we have a stratified random sample.

The study was therefore designed to investigate the efficiency of Neyman allocation procedure over equal and proportional allocations in a stratified random sampling.

The data used for this research were primary data collected from ten Markets in Abeokuta, Ogun State, Nigeria on the prices of Peak Milk. A stratified random sampling scheme was used in selecting sample of size 350 from 10 markets. Each market stands as a stratum and independent sample was selected randomly based on equal, proportional and Neyman allocations and statistic was obtained from each stratum and combined estimate of the separate statistic was also obtained for each of the allocation procedure.

The mean and variance under Neyman allocation procedure were 1356.672 and 21.45 respectively. For proportional allocation, the mean was 1349 and variance was 38.98 while equal allocation gave mean to be 1352 and variance was 170.3238. Neyman/Optimum allocation procedure gave the least variance. This was followed by proportional allocation and equal allocation. Neyman allocation procedure is the best selection procedure. Hence, to obtain estimate of high precision for the prices of Peak Milk in the markets, Neyman allocation procedure is the most efficient.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Statistics in BAE Systems Munitions

Robert Langford
BAE Systems Munitions, Glascoed, UK

Munitions designs, develops and manufactures a comprehensive range of munitions for both the MOD and other export countries. Our main customer is the UK MOD and we currently supply +80% of their conventional munitions. The munitions we produce range from small arms ammunition to mortar bombs and artillery shells. Munitions' head office is located at Glascoed (South Wales) but there are two further manufacturing sites located at Radway Green (Cheshire) and Washington (Tyne and Wear) and two testing facilities at Ridsdale (Northumberland) and Bishopton (Stratchclyde).

Poster viewing B

Wednesday 4 September 1.30pm - 2.30pm

Development and validation of a model for longitudinal survey optimum response

O.M. Olayiwola[1], A.A Okunlade[2], S.A. Mustapha[3]
[1]Department Of Statistics, College Of Natural Sciences, Federal University Of Agriculture, Abeokuta, Ogun State, Nigeria, [2]Nigerian Institute Of Science Laboratory Technology, Ibadan, Nigeria, [3]C.R.C, Oyo, Nigeria

Non-response in surveys is an indicator of data quality; they introduce bias in the estimates. Previous studies examined the effects of response predictors on response rates. There is a dearth of information about models which focus on the interaction effects of response predictors on response rates. The study was designed to develop and validate a model which reduce non-response and achieve optimum response.

A two-stage stratified random sampling scheme was used in selecting 750 households. Households were interviewed in five waves to collect data on demographic characteristics and response predictors. Multi-way contingency tables were constructed. A model was developed by breaking the predictors of response into levels and their interaction effects were introduced into Hawkes and Plewis model. A log-linear model was fitted to construct contingency tables. A Response Surface Polynomial Model (RSPM) was constructed and subjected to canonical analysis.

The respondents' mean age and modal family size were 51.86.9 and 3 respectively, 64.8% were females, 52.8% were Muslims and majority (88.9%) were employed. Response rates for waves were 72.7%, 82.0%, 81.3%, 82.0% and 80.7% respectively. Patterns of participation in the surveys were: continued participation (80.0%), monotone attrition (2.0%), new entry (8.0%), occasional non-response (3.3%) and occasional-response (6.7%). The developed model showed that family size duration of interview and their interaction significantly ($p < 0.05$) determined response rate. The response was optimum when the family size was three and duration of interview was twelve minutes.

Family size, duration of interview and their mutual interaction play key roles in obtaining optimum response.

Poster viewing B

Wednesday 4 September 1.30pm - 2.30pm

Bayesian bandwidth estimation for a nonparametric functional regression model with mixed types of data and unknown error density

Han Lin Shang
University of Southampton, Southampton, UK

We investigate the issue of bandwidth estimation in a nonparametric functional regression model with function-valued, continuous real-valued and discrete-valued regressors under the framework of unknown error density. We propose to approximate the unknown error density by a location-mixture of Gaussian densities with means being the individual residuals, and variance a constant parameter. This proposed mixture error density has a form of a kernel density estimator of residuals, where the regression function is estimated by the functional Nadaraya-Watson estimator that admits mixed types of regressors. We put forward a Bayesian bandwidth estimation approach that can simultaneously estimate the bandwidths in the kernel-form error density and functional Nadaraya-Watson estimator. We derive a kernel likelihood and posterior density for the bandwidth parameters under the kernel-form error density. A series of simulation studies demonstrated the estimation accuracy of the regression function and error density for the proposed Bayesian approach. Illustrated by the spectroscopy data, we applied the proposed Bayesian procedure to a nonparametric functional regression model with mixed type of regressors, and we found that it has a more superior point forecast accuracy than several existing functional regression models.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Semi-parametric predictive inference for bivariate data using copulas

Noryanti Muhammad, Frank Coolen, Tahani Coolen-Maturi
University of Durham, Durham, UK

Nonparametric Predictive Inference (NPI) is a frequentist statistical framework for inference on a future observation based on past data observations. NPI uses lower and upper probabilities to quantify uncertainty based on only few modelling assumptions. NPI has been presented for Bernoulli data, ordinal data, multinomial data and real-valued data, possibly with right-censored observations.

In this research, we develop NPI for bivariate data with the use of copulas. A copula is a joint distribution function whose marginals are all uniformly distributed on [0,1] and it can be used to model the dependence separately from the marginal distributions. Initially, we use NPI for the marginals combined with a parametric copula, leading to a new semi-parametric method.

We investigate the performance of this method by presenting results for different parametric copulas combined with several estimation methods. Some applications are presented to illustrate the proposed method, and we outline related research challenges.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

A penalised quality index for identifying number of clusters

Marica Manisera, Marika Vezzoli
University of Brescia, Brescia, Italy

Identifying number of clusters is a compulsory step in cluster analysis. Many papers investigated this issue and several criteria have been proposed in the literature. The objective of this study is to propose an R-routine that implements a new method able to automatically identify the optimal number of clusters in a hierarchical clustering algorithm.

Starting from the idea of pruning, introduced in the framework of Regression Trees, we propose to use a penalised internal cluster quality index in order to identify the best cut in the dendrogram. It is composed by an index able to validate clustering and a penalty depending on the number of groups. The procedure aims to stop the unsupervised clustering algorithm that partitions the variables' space identifying a reduced, therefore interpretable, number of clusters.

This study results in an R routine that automatically identifies the best number of groups and replaces the discretional choice usually required to cut a dendrogram. In detail, the procedure adapts the cv.tree package in R, widely used in the Regression Trees framework, and provides a sequence of subtrees associated to all the possible cuts in the dendrogram. Each subtree in the sequence is characterized by its number of groups, the value of the penalised internal cluster quality index, and the optimal value of the penalty parameter, which is obtained algorithmically by means of cross-validation. The proposed procedure can also be applied to Dynamic Cut Tree. In this paper, we tested our procedure on simulated data with known structure.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Consequences of ignoring clustering in linear regression

Georgia Ntani, David Coggon, Hazel Inskip
MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK

Objectives

It is widely stated that ignoring clustering of data produces unbiased but spuriously precise estimates. We aimed to explore differences in regression coefficients and the ratios of the associated Standard Errors (SEs) when derived from random-intercept Multi-Level (ML) models as compared with Ordinary Least Squares (OLS) models that ignore clustering of data.

Methods

Hierarchical datasets (70 clusters with 100 observations per cluster) were simulated with varying degrees of within cluster dependency of observations (characterised by intraclass correlation coefficient (ICC)) and dispersion of cluster-specific mean values for the independent variable (sd(x_bar)). The underlying regression coefficient was arbitrarily set to one as were the within cluster standard deviation of x and standard deviation of y given x. For each simulated dataset, the parameters of interest from the ML and OLS models were estimated and compared.

Results

When ICC was small (5%) regression coefficients estimated by the two methods were nearly identical, but SEs from ML modelling were greater, reaching a maximum ratio of ~2.4 as sd(x_bar) increased. When ICC was large (75%), regression coefficients differed by -0.22 to 0.25. For large sd(x_bar), SEs from ML modelling were again higher, but for small sd(x_bar) (<1.7), SEs were lower with OLS than with ML modelling (ratios as low as 0.5).

Conclusions

The common belief that application of OLS regression to clustered data gives the same regression coefficient as ML modelling but with smaller SE holds when ICC is low. However, when ICC is high and sd(x_bar) is small that is not necessarily correct.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Local community awareness and use of wetland resources in Kyeizooba sub-county Bushenyi district – Uganda

Dorothy Mubuuke
Makerere University, Kampala, Uganda

The study was aimed at finding out the people's level of awareness and use of wetland resources in Kyeizooba Sub County, in Bushenyi district. The researcher used a descriptive survey to get opinions of the community on awareness programs and wetland uses. Questionnaires, interview guides and observation guides were also used. Data collected were computed using the chi-square test and percentages

The study established that local communities living in and around Kyeizooba wetlands were aware that wetland degradation is the major problem in the area. There are awareness and education programs being conducted in the area. The field staff educated the local communities on the conservation of the environment and wetland resources in the Sub County and Bushenyi district in general.

Whereas the ongoing awareness-raising programs have had a great impact on the environmental awareness of the local communities, their impact on wetland conservation practices carried out by the local communities has been low, with many people draining wetlands.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Sufficiency and adequacy of the t-ratio in determining the presence of multicollinearity in a regression model sufficiently

Adebayo Agunbiade
Olabisi Onabanjo Univeristy, Ago Iwoye, Ogun State, Nigeria

One of the violations of the assumptions made regarding the Classical Linear Regression Model is Multicollinearity. This phenomenon is one of the most misunderstood problems in multiple regression analysis. The cause, effects and consequences of multicollinearity are well documented in the literature. The focus of this research is in the detection of multicollinearity, which deals with the problem of intercorrelation of explanatory variables. This paper therefore investigates the sufficiency and adequacy of the t-ratios only to confirm the presence of multicollinearity problem in a multiple regression equation. To achieve this, a three-equation simultaneous model with three multicollinear exogenous variables is presented. Monte Carlo simulation indicates that the asymptotic results provide a better estimate with Variance Inflation Factor. It was established that a combination of one or two of the criteria considered will suffice and not just the t-ratio only in determining the presence of multicollinearity.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Quality of midwifery care in Uganda – case study Soroti district

Resty Nansubuga
Makerere University, Institute of Statistics and Applied Economics, Kampala, Uganda

A midwife is a person trained to provide care to women during pregnancy, labour and after delivery; trained to recognise problems. This study determined the quality of care provided by midwives in Soroti District, Uganda. Specifically, it identified training needs, gaps in knowledge and other barriers to accessibility of emergency obstetric care services in the area. A cross-sectional descriptive study was designed to assess two hospitals, two health centres, and four sub-dispensaries over a two-month period. Data were collected from patients and attendants, midwives, pregnant women exiting from antenatal clinics, and health unit records. The main outcome measures were the ability of midwives to provide antenatal care at a minimum standard, or recall causes, signs, symptoms or the management of common obstetric complications; and patients' and patient attendants' perceptions of care received. Overall, the antenatal care offered was routine, with no emphasis on women at risk. Many midwives were providing poor quality antenatal and delivery care due to their inability to identify and manage women with, or at risk of, pregnancy complications. The main factors responsible for the poor quality of care were inadequate pre-service and in-service training, lack of technical support supervision and absence of standard treatment guidelines.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Sub-state life expectancy estimation using the methodology for small population

Prasanta Barman[1], Labananda Choudhury[2]
[1]Kohima Science College, Jotsoma, Kohima, Nagaland, India, [2]Hauhati University, Guwahati, Assam, India

Introduction

Life expectancy at birth is one of the most preferred indicators in demographic and health analysis. Though the office of the Registrar General of India periodically publishes the life expectancies for the bigger states, for smaller states no figures are available. Nagaland, the sixteenth state of the Indian union with its population 1,980,602 is one of the smallest tribal states of India.

Objectives

In this study, an attempt has been made to estimate the life expectancy at birth along with its standard error for Kohima and Dimapur districts of Nagaland during 2010-2011. Also, to investigate the shape of the sampling distribution of the estimated life expectancy at births.

Materials and Methods

The data used is primary one collected through a household survey conducted in both the districts. Life expectancy at birth is estimated applying Silcocks methodology of small area life expectancy estimation. Additionally, a Monte Carlo Simulation technique is also used to generate the sampling distribution of life expectancy at birth, which is expected to provide a check on the result of the Silcocks formula and indicated the shape of the distribution.

Results

The result indicates that the life expectancy at birth for males and females is 66.17 and 67.84 years respectively. The mean life expectancy at birth is observed as 66.20 years for males and 67.40 years for females. Life expectancy at birth has an approximately normal distribution and the simulated result approximated very well to the Silcocks methodology of life expectancy estimation.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

A Model of Neonatal Mortality in Republika Srpska

Rainhard Bengez[1,2]
[1]TU Muenchen, Muenchen, Germany, [2]National University of Taipei, Taipei, Taiwan

Background

Social paediatrics is an interdisciplinary and empirical science investigating the frequency and distributions of disturbance of health and developmental disorder of infants, children, teenagers, and young adults related to their family situation and social or natural environment. Neonatal mortality depends on the age of gestation and is a good indicator for early childhood prevention, socio-economic development and quality of medicine and life. These rates vary across regions and districts.

Objectives

In this contribution we analyse neonatal mortality in Republika Srpska, Bosnia, with different spatial survival models. The application is based on a not yet published study conducted between 1996 and 2010.

Models and Methods

This class of models allows to measure small spatial effects simultaneously with possibly nonlinear or time-dependent factors. For the analysis we are using Bayesian Inference and MMC-simulation (Markov chain Monte Carlo). The methods discussed in this work are not based on parametric models, but use different temporal and spatial resolution.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Evaluation of an enhanced model of partnership working to support universal speech and language provision in the early years

Kate Jones[2,1], Thomas King[1], Helen Stringer[1]
[1]Newcastle University, Newcastle upon Tyne, UK, [2]Northumberland Church of England Academy, Northumberland, UK

Language develops naturally in children but it is also one of the principal goals for education in the early years as it is a precursor to literacy. Oral communication is also fundamental to socialisation and is being recognised as having continuing significance for older children and into the workplace, beyond its underpinning of reading and writing. A universal intervention characterised by training for teachers in creating communication friendly environments and targeted small group activity for children with specific deficits was supported by specialist intervention. Training was propagated from the original school to the other schools of the academy to reach a total of 409 children over a period of two years. The intervention was monitored by termly language assessments in five domains by specialists yielding an average of five observations of each measure per child.

A longitudinal mixed effects model with linear random effects was used to evaluate the effectiveness of the intervention regime and shows substantial effects. While in a number of domains this is shown in weaker children being accelerated towards the test ceiling, the British Picture Vocabulary Scale test score has no ceiling for the age range tested (3-6 years). The intervention has a substantial effect of both catching up and general improvement although the precise effect is complicated by the endogeneity of treatment assignment (universal, targeted, specialist) to the outcome measure. The trajectory model is helpful in providing an intuitive model of development and individual level representations of progress and variability.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Stochastic Hybrid Modelling of the cell decisions leading to bone formation

Rodrigo Assar[1,2], Martín A. Montecino[1,3]
[1]Fondecyt 3130762, Centro de Investigaciones Biomédicas, Facultad de Ciencias Biológicas and Facultad de Medicina, Universidad Andrés Bello, Santiago, Chile, [2]Fundación INRIA Chile, Santiago, Chile, [3]Fondap 15090007, Center for Genome Regulation, Santiago, Chile

Biological system behaviours are consequence of interacting processes, which are affected by external factors often not controlled. These processes connect different hierarchy levels, and it is common to observe that changes in the conditions (such as external signals) modify the behaviour of the systems. We need a modelling approach being able to integrate all these characteristics to simulate in particular the bone formation process. Understanding this process could allow us to anticipate dysfunctions such as osteoporosis, and looking for more effective and less side-effect-filled treatments. We considered a Hybrid Systems approach. The system variables evolve according to continuous models, but deterministic, stochastic and non-deterministic mode transitions can change the definition of these models. These mode transitions respond to system behaviour law changes and are described by Stochastic Transition Systems. Hybrid models allow us to combine different types of dynamics and build composed and hierarchical models. One uses continuous dynamics for gradual changes, deterministic for complete knowledge, and stochastic or non-deterministic to describe behaviours with imprecise or incomplete information. We obtained a hybrid description of the dynamics of bone precursor cells, osteoblasts and adipocytes. The concentration of each cell lineage is affected by mode signals, such as the activation of the Wnt pathway favouring the osteoblast lineage, as well as apoptosis stimuli and stochastic factors changing the system conditions. Epigenetic inheritance is described by stochastic modelling, and intermediate states of predisposition to differentiate into specific lineages. Our model allows us to predict and simulate bone formation responses to signal stimuli.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

How robust is the Linear Logistic Test Model when the Q-matrix is misspecified?

George MacDonald, Jeffrey Kromrey, Yi-Hsin Chen
University of South Florida, Tampa/Fl, USA

Marginal maximum likelihood was used to estimate Linear Logistic Test Model (LLTM) cognitive components and item difficulty. Empirical Bayes was used to estimate person ability. The factors manipulated in the simulation study included overall type of misspecification (under-specification, balanced misspecification, over-specification), per cent misspecification (1%, 5%, 10%, 15%), sample size (20, 40, 80, 160, 320, 640, 1280), Q-matrix density (64%, 46%), skewness of person distribution (-0.5, 0, 0.5), and test length (20, 40, 60). Statistical bias, RMSE, CI coverage, and CI width were computed across the set of replications for cognitive components and person ability. Statistical bias and RMSE were computed for LLTM item difficulty. As the Q-matrix moves away from the truth cognitive components, and item difficulty estimates become progressively more biased. Under-specification results in estimates that have a positive bias. Balanced misspecification and over-specification of the Q-matrix result in cognitive components estimates and item difficulty estimates that are progressively more negatively biased. Results were interpreted for the design factors and recommendations provided.

Poster viewing B
Wednesday 4 September 1.30pm - 2.30pm

Dynamic process monitoring using Partial Least Squares and extensions

Bothinah Altaf, Gary Montague, Elaine Martin
Newcastle University, Newcastle upon Tyne, UK

Multivariate Statistical Process Monitoring (MSPM) plays an important role in competitive industrial world. It helps to ensures satisfactory performance and production in term of safety and quality. By detecting the abnormal events, one can eliminate the root causes and hence, improves process efficiency and quality products. As most of industrial processes are generally dynamic in nature, the traditional MSPM methods such Partial Least Squares (PLS), which assume steady state relationships, show unsatisfactory performance. As they do not consider the dynamic relationships in the industrial processes, those increase the false alarm rate and decrease the effectiveness of the monitoring system. Therefore, a dynamic version of the MSPM method was developed.
The rapid development of automated data acquisition systems increases the number of measured variables. To improve the interpretability of the process behaviour a Multi-Block Dynamic PLS (MBDPLS) is proposed to monitor the whole process and the individual unit operation. It is extended to Adaptive Multi-Block Dynamic PLS (AMBDPLS) for process monitoring in real time basis, where the main challenge is the impact of outlying samples as the model will incorrectly adapt to these samples. Therefore, robust AMBDPLS is proposed, which is resistance to outlying samples. The proposed methods were tested and compared using the data generated from a simulation of Tennessee Eastman Process. Application of the proposed method showed a superior performance compared to the MBDPLS and AMBDPLS. It decreases the false alarm rate and increases the fault detection rate compared to MBDPLS and AMBDPLS respectively.

Author index

A

Abdelsalam · 153
Agarwal · 73
Agunbiade · 169
Aitken · 17, 30
Alexandrov · 124
Altaf · 176
Altman · 61
Anastasiou · 156
Ansell · 33
Ariyibi · 53
Asimit · 109
Askew · 112, 149
Assar · 174

B

Barber · 108
Barman · 171
Barnes · 129
Barrett · 46
Bengez · 172
Bertsch McGrayne · 101
Betancourt · 66
Bird · 83, 138
Browne · 40
Buck · 126
Buckland · 68
Burke · 27

C

Cheshire · 119
Chigbu · 146
Choudhury · 142, 151
Cole · 62
Collingwood · 3
Coolen-Maturi · 32
Corradi · 25
Cox · 21, 83

D

Daneshkhah · 90
Daniels · 21
Das · 147, 148
Demsar · 117
Dent · 135
Dick · 70
Dietz · 10
Dregan · 76
Duembgen · 121

E

Edmondson-Jones · 139, 141
Ensor · 29
Eslava-Gomez · 145

F

Farrow · 51
Figini · 34
Finazzi · 130
Finke · 103
Fry · 24

G

Galwey · 48
Geraci · 63
Ghee · 72
Goldstein · 77
Grant · 60
Greenaway · 116
Gunsoy · 54

H

Han, Yang · 5, 155
Harper · 57
Harron · 20
Heaps · 79
Henmi · 144
Hobert · 22
Hofmeyr · 104
Hunt · 111

I

Iniesta-Martinez · 21

J

Jackson · 19
James · 115
Johnson · 93
Jonathan · 137
Jones, Chris · 64
Jones, Paul · 97

K

Kamenos · 128
Kavanagh · 36
Kelly · 7
Kimber · 50
King · 55, 173
Kontopantelis · 16, 150
Korkas · 8
Koutoumanou · 87
Kunst · 6

L

Langford · 162
Larcombe · 157
Lau · 52
Leckie · 14
Lee · 65
Lewis · 133
Lin · 58
Lynch · 122

M

MacDonald · 175
Main · 82
Manca · 114
Manisera · 59
Mason · 92
Matejic · 129
Maugis · 15
McHale · 132
Mehrhoff · 83
Moreno · 23
Moriarty · 105
Morris · 91
Moschandreas · 140
Mubuuke · 168
Muhammad · 165
Muirhead · 100
Mukhaiyar · 35
Mylne · 80

N

Nansubuga · 170
Nason · 136
Naumann · 39
Naylor · 71
Newell · 96
Nielsen · 120
Noufaily · 85
Ntani · 167

O

Olatayo · 143
Olayiwola · 161, 163

P

Pankhurst · 99
Parnell · 127
Petersen · 74
Philipson · 42
Potts · 95
Powell · 113

R

Rasulo · 56
Ray · 106
Reeves · 159
Relton · 134
Revie · 43
Riley · 28
Roberts · 18
Rogers · 45
Rougier · 81

S

Sangalli · 107, 131
Sarma · 152
Scott · 69
Shang · 98, 164
Shimadzu · 94
Singh Roy · 44
Smith, Michael · 88
Smith, Mike · 123
Smith, Ron · 67
Soberon Velez · 31
Spencer · 41
Sperrin · 38
Springate · 75
Stevenson · 102

T

Tampubolon · 13
ter Braak · 118
Tinsley · 72
Tinsley, Matthew · 129
Touloumis · 49
Trela-Larsen · 12

V

Vallejos · 86
Veit-Wilson · 129

Vernon · 78
Vezzoli · 166

W

Walter · 47
Wang · 160
Waruszynski · 4
Wedge · 125
Wei · 26
White · 37
Wilke · 89

Y

Yarkiner · 11

Z

Zanetti Chini · 158
Ziel · 9
Zwick · 2