ISCB 2015, Utrecht, the Netherlands
Invited session

STRengthening Analytical Thinking for Observational Studies (STRATOS)

Wednesday, August 26

9.00 - 11.00

Katherine Lee - University of Melbourne, Australia
Regression modelling with missing data: principles, methods, software and examples

Missing data are ubiquitous in research, and are often an unwelcome headache for hard-pressed analysts. Depending on the context, and the reasons for the missing data, the simple solution of analyzing the subset of complete records may or may not be adequate. Therefore, focusing on regression modelling, the Missing Data Topic Group has set out to review the issues raised by missing data, outline when a complete records analysis is likely to be 'good enough'. Then, for situations when a more sophisticated approach is needed, we review, illustrate and critique the established methodology and associated software. Using data from the Youth Cohort Study of England and Wales (a publically available dataset) we illustrate how a masters-level analyst faced with a missing data problem should engage with our proposed materials, and in particular what they should learn and put into practice

Stephen Evans - London School of Hygiene & Tropical Medicine, London, United Kingdom
STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative: study design

Appropriate and valid study design is crucial for valid conduct of observational studies(OS). These contribute to establishing causal effects, together with other evidence (e.g. mechanistic studies, clinical trials), though some OS do not attempt to infer causality, e.g. prognostic studies or estimation of. The appropriateness of any design depends on the research question, in the context of current theory and knowledge, availability of valid measurement tools, and the proposed uses of the results. In theory, some study designs are seen as less biased; in practice validity is topic- and context-specific. Hierarchies of OS are often proposed, e.g. cohort studies high, followed by case-control and cross-sectional studies. However, their relative validity represents a continuum, and 'less valid' study designs may yield equally valuable information. It is unusual for a single OS to deliver definitive results, so assessing epidemiologic evidence almost always involves combining information from different study populations, designs, investigators, and methods. None can be perfect; rather, the aim is to contribute to the pool of knowledge for a particular issue, in a particular population and risk period. Valid design involves a context-specific balance between these competing considerations. For example a blood test may yield better estimates of exposure (reducing information bias), but reluctance to give blood may increase missing data (potentially increasing information bias and increasing random error) and lower response rates (increasing selection bias). Whatever design is used, it is important to be able to conduct sensitivity analyses, and/or including control exposures expected to have null effects.

Aris Perperoglou - University of Essex, United Kingdom
Multivariable regression modelling using splines. A review of available packages in R

Building explanatory models depends heavily on two interrelated aspects: the selection of variables and their functional form. To deal with complex functional forms of continuous variables statisticians have developed a variety of spline methods, many of which are implemented in R packages. This project will investigate options that are available in R for building multivariable regression models with splines, compare between different approaches and provide practical guidance on available software. Out of a total of more than 6200 packages on CRAN (May 2015) we identified a subset of approximately 100 packages that have some spline related function. Packages were classified into two categories, the ones that create spline bases and those that fit regression. For the first type of packages we identified the types of splines available, whether a user can define degrees of freedom, number and position of knots and if there are available methods for determining the smoothing parameters. For regression packages we looked into types of regression models, whether the package includes criteria for the significance of a non-linear effect or graphical tools to identify variables that should be transformed, procedures for variable selection and multivariable methods. In this presentation we will present our first findings. We will show what are the most used packages, what are their interdependencies and their basic features. Furthermore, we will discuss the framework for further research, where we will evaluate the quality and performance of packages with the aim to provide detailed guidelines for applied researchers.

Terry Therneau - Mayo Clinic, Rochester, Minnesota, United States
The STRATOS survival task group

The survival task group has only recently organized. Since any of the important topics in modeleling survival data overlap with those of other task groups (selection of variables and functional form, measurement error, causal effects, etc.) our intial work will focus on aspects that are particular to survival data. These include, in no particular order - recurrent and competing events, and multi-state models - time dependent effects - time varying covariates and the common error of conditioning on the future, e.g., survival curves of responders vs non-responders - relative uses and merits of the proportional hazards, additive hazards, and accelerated failure time models - relative survival - interval censoring - random effects ("frailty") - joint modeling of survival and longitudinal markers This talk will give an overview of the topics and our directions along with a deeper look at issues in topic, and hopefully engender wider discussion and/or participation in the project.