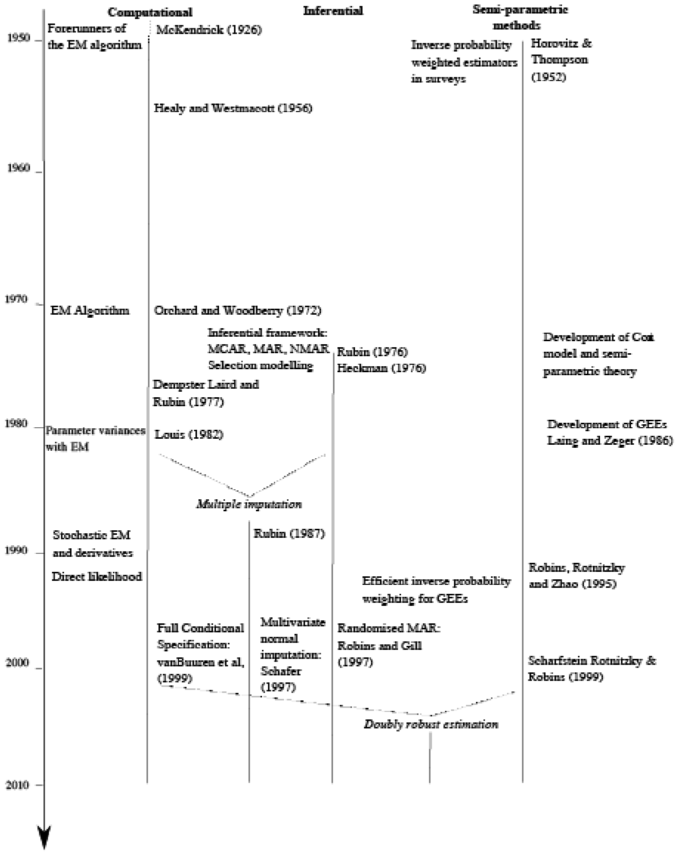
STRengthening Analytical Thinking for Observational Studies (STRATOS): Introducing the Missing Data topic group (TG1)


Computational
Inferential
Semi-parametric methods
1950
1960
1970
1980
1990
2000
2010
Forerunners of the EM algorithm
McKendrick (1926)
Inverse probability weighted estimators in surveys
Horovitz & Thompson (1952)
Healy and Westmacott (1956)
EM Algorithm
Orchard and Woodberry (1972)
Inferential framework: MCAR, MAR, NMAR Selection modelling
Rubin (1976)
Heckman (1976)
Development of Cox model and semi-parametric theory
Dempster Laird and Rubin (1977)
Parameter variances with EM
Louis (1982)
Development of GEEs Laing and Zeger (1986)
Multiple imputation
Stochastic EM and derivatives
Rubin (1987)
Direct likelihood
Efficient inverse probability weighting for GEEs
Robins, Rotnitzky and Zhao (1995)
Full Conditional Specification: vanBuuren et al, (1999)
Multivariate normal imputation: Schafer (1997)
Randomised MAR: Robins and Gill (1997)
Scharfstein Rotnitzky & Robins (1999)
Doubly robust estimation

James Carpenter (1) and Katherine J Lee (2) on behalf of STRATOS TG1.

(1) London School of Hygiene & Tropical Medicine and MRC Clinical Trials Unit at UCL. Email: james.carpenter@lshtm.ac.uk

(2) Murdoch Children’s Research Institute, Melbourne and Department of Paediatrics, University of Melbourne. Email: Katherine.lee@mcri.edu.au.

In the last issue of the Bulletin, the aims and recent activities of the STRATOS initiative were described. This initiative is divided into a number of topic groups, which provide a focus for specific research methodologies. Here, we introduce the Missing Data topic group.

Missing data are ubiquitous in medical and social research, and there is a long history of methods for coping with the difficulties they raise. As the Figure illustrates, early work (dating back at least to 1926) was largely computational, addressing the question of how to perform the analysis when data are missing. Later, the focus switched to establishing a valid inferential framework when data are missing, giving rise to an array of methods for handling missing data including multiple imputation, inverse probability weighting and Bayesian methods. This research has also linked with work in non-parametric statistics and gave rise to doubly robust estimation (which seeks to combine the best of imputation and inverse probability weighting). How best to handle missing data remains an active area of research, which has recently been brought together in the Handbook of Missing Data Methodology (Molenberghs et al, 2015).

Most researchers, whether or not they have a formal statistical training, are aware of the challenges missing data raise, and at least some of the approaches available. In particular, multiple imputation has become increasingly popular for handling missing data, fuelled by the development flexible software that is now available in all the leading statistical software packages.

However, in our experience, many researchers remain unclear about the practical value of more sophisticated approaches like multiple imputation compared with restricting the analysis to those with no missing data. Often, they may also be unsure about how methods like multiple imputation, full information maximum likelihood, the Expectation-Maximization algorithm and inverse probability weighting relate to each other, and the practical implications of this in their specific setting.

The Missing Data topic group is working to address these issues through a series of linked papers. Following the aims of the initiative (Sauerbrei et al,2014(we seek to address three audiences: those engaged in quantitative research, but without a formal statistical training (level 1); those with a statistics training to masters level (level 2), and those interested or active in missing data research (level 3).

The topic group consists of researchers known internationally for their work across the broad spectrum of theoretical and practical methodology for missing data: Rod Little (Michigan, USA), Andrea Rotnitzky (Harvard, USA), Joe Hogan (Brown, USA) Els Goetghebeur (Gent, Belgium), Ian White (UCL, London), Kate Tilling (Bristol, UK) and Melanie Bell (Arizona, USA). James Carpenter (London, UK) and Katherine Lee (Melbourne, Australia) chair the group.

The topic group played an active role in the 2016 STRATOS workshop at the mathematical sciences research centre in Banff, Canada, where James gave an overview lecture ‘Handling missing data in observational studies: challenges for teaching and research’, which is publicly available at www.stratos-initiative.org/node/49. Building on this, currently the Topic Group has three papers close to submission, which will be highlighted on the STRATOS website when they are available (http://stratos-initiative.org).

The first paper, led by Rod Little, compares three popular methods for handling missing data in a social science setting: complete cases, weighting and multiple imputation. This paper is aimed at researchers with relatively little formal statistical training, and uses a simple example from the UK’s Youth Cohort Study, a publically available dataset, to build intuition for how biases may be caused by missing data, and the pros and cons of these approaches. It includes practical guidance on which methods are preferable in which situations.

The second paper, led by Katherine Lee, targeted at those using observational data for medical research, is likewise aimed at researchers with relatively little formal statistical training. Building on related work in clinical trials (which is primarily concerned with missing outcome data), this paper aims to provide and illustrate a practical framework for the analysis of partially observed data and subsequent reporting. Again this paper will include a worked example from the Youth Cohort Study along with example code.

The third paper addresses statisticians, and discusses both theoretically and with examples, the utility of key approaches to the analysis of partially observed data, in particular: full Bayesian analysis, multiple imputation, inverse probability weighting, doubly robust estimation, direct maximum likelihood and the EM algorithm. This paper will include the statistical code for each of the approaches and will be based on an example from a publically available dataset so that readers will be able to re-create the analyses presented. The aim in this paper is to providing the necessary tools for researchers wishing to apply these approaches in practice.

As missing data, and in particular the most appropriate way to handle the issues it raises, varies depending on the setting and the statistical models required for analysis, TG1 is also linked to other groups in the STRATOS initiative, particularly the Initial Data Analysis (STRATOS Topic Group 3), Measurement Error Misclassification (STRATOS Topic Group 4) and Causal Inference (STRATOS Topic Group 7).

The Missing Data Topic Group is keen to stimulate interactions with other researchers; James and Katherine welcome any comments and suggestions on its work, future issues that would be useful to consider and illustrative datasets.

References in text

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S. and Carpenter J R on behalf of the STRATOS initiative (2014). STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Statist Med*, **33**, 5413-5432.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis A. and Verbeke, G (2015). *Handbook of Missing Data Methodology*. New York: CRC press

References for Figure

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,* **39**, 1-38.

Healy, M. J. R. and Westmacott, M. (1956) Missing values in experiments analyzed on automatic computers. *Applied Statistics*, **5,** 203-206.

Heckman, J. J. (1976) THe common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement,* **5**, 475-492

Horvitz, D. G. and Thompson, D. J. (1952) A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* **47**, 663-685.

Laing, K-Y and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models.

Biometrika, **73**, 13-22.

Louis, T. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B*, **44**, 226-233

McKendrick, A. G. (1926) Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society,* **44**, 98-130.

Orchard, T. and Woodbury, M. (1972) A missing information principle: theory and applications. *Proceedings of the Sizth Berkely Symposium on Mathematics, Statistics and Probability, editors Le Cam, L. M., Neyman J. and Scott, E. L,* **1**, 697-715

Robins, J. M. and Gill, R. (1997) Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, **16**, 39-56.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106-121.

Rubin, D. B. (1976) Inference with missing data. *Biometrika*, **63**, 581-592.

Rubin, D. B. (1987) *Multiple imputation for non-response in surveys.* New York: Wiley

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999) Adjusting for non-ignorable drop-out using semiparametric nonresponse models.

Schafer, J. L. (1997) *Analysis of incomplete multivariate data*, London: Chapman and Hall

van Buuren, S. and Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681-694.