

Multiple Imputation and its Application

JAMES R. CARPENTER AND MICHAEL G. KENWARD

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK.

A practical guide to analysing partially observed data.

Collecting, analysing and drawing inferences from data is central to research in the medical and social sciences. Unfortunately, it is rarely possible to collect all the intended data. The literature on inference from the resulting incomplete data is now huge, and continues to grow both as methods are developed for large and complex data structures, and as increasing computer power and suitable software enable researchers to apply these methods.

This book focuses on a particular statistical method for analysing and drawing inferences from incomplete data, called Multiple Imputation (MI). MI is attractive because it is both practical and widely applicable. The authors' aim is to clarify the issues raised by missing data, describing the rationale for MI, the relationship between the various imputation models and associated algorithms and its application to increasingly complex data structures.

Multiple Imputation and its Application:

- Discusses the issues raised by the analysis of partially observed data, and the assumptions on which analyses rest.
- Presents a practical guide to the issues to consider when analysing incomplete data from both observational studies and randomized trials.
- Provides a detailed discussion of the practical use of MI with real-world examples drawn from medical and social statistics.
- Explores handling non-linear relationships and interactions with multiple imputation, survival analysis, multilevel multiple imputation, sensitivity analysis via multiple imputation, using non-response weights with multiple imputation and doubly robust multiple imputation.
- Is supported by a supplementary website www.wiley.com/go/multiple_imputation featuring datasets and illustrative code, with the freely available REALCOM impute software as well as SAS, Stata, MLwiN and R.

Multiple Imputation and its Application is aimed at quantitative researchers and students in the medical and social sciences with the aim of clarifying the issues raised by the analysis of incomplete data, outlining the rationale for MI and describing how to consider and address the issues that arise in its application.

STATISTICS IN PRACTICE

A series of practical books outlining the use of statistical techniques in a wide range of applications areas:

- HUMAN AND BIOLOGICAL SCIENCES
- EARTH AND ENVIRONMENTAL SCIENCES
- INDUSTRY, COMMERCE AND FINANCE



Also available
as an e-book

Cover photograph courtesy of Harvey Goldstein

Cover design by Cylinder



WILEY

wiley.com

ISBN 978-0-470-74052-1



9 780470 740521

Contents

Preface	xi
Data acknowledgements	xiii
Acknowledgements	xv
Glossary	xvii
 PART I FOUNDATIONS	 1
1 Introduction	3
1.1 Reasons for missing data	4
1.2 Examples	6
1.3 Patterns of missing data	7
1.3.1 Consequences of missing data	9
1.4 Inferential framework and notation	10
1.4.1 Missing Completely At Random (MCAR)	11
1.4.2 Missing At Random (MAR)	12
1.4.3 Missing Not At Random (MNAR)	17
1.4.4 Ignorability	21
1.5 Using observed data to inform assumptions about the missingness mechanism	21
1.6 Implications of missing data mechanisms for regression analyses	24
1.6.1 Partially observed response	24
1.6.2 Missing covariates	28
1.6.3 Missing covariates and response	30
1.6.4 Subtle issues I: The odds ratio	30
1.6.5 Implication for linear regression	32
1.6.6 Subtle issues II: Subsample ignorability	33
1.6.7 Summary: When restricting to complete records is valid	34
1.7 Summary	35
 2 The multiple imputation procedure and its justification	 37
2.1 Introduction	37
2.2 Intuitive outline of the MI procedure	38

2.3	The generic MI procedure	44
2.4	Bayesian justification of MI	46
2.5	Frequentist inference	48
2.5.1	Large number of imputations	49
2.5.2	Small number of imputations	49
2.6	Choosing the number of imputations	54
2.7	Some simple examples	55
2.8	MI in more general settings	62
2.8.1	Survey sample settings	70
2.9	Constructing congenial imputation models	70
2.10	Practical considerations for choosing imputation models	71
2.11	Discussion	73

PART II MULTIPLE IMPUTATION FOR CROSS SECTIONAL DATA **75**

3	Multiple imputation of quantitative data	77
3.1	Regression imputation with a monotone missingness pattern	77
3.1.1	MAR mechanisms consistent with a monotone pattern	79
3.1.2	Justification	81
3.2	Joint modelling	81
3.2.1	Fitting the imputation model	82
3.3	Full conditional specification	85
3.3.1	Justification	86
3.4	Full conditional specification versus joint modelling	87
3.5	Software for multivariate normal imputation	88
3.6	Discussion	88
4	Multiple imputation of binary and ordinal data	90
4.1	Sequential imputation with monotone missingness pattern	90
4.2	Joint modelling with the multivariate normal distribution	92
4.3	Modelling binary data using latent normal variables	94
4.3.1	Latent normal model for ordinal data	98
4.4	General location model	103
4.5	Full conditional specification	103
4.5.1	Justification	103
4.6	Issues with over-fitting	104
4.7	Pros and cons of the various approaches	109
4.8	Software	110
4.9	Discussion	111
5	Multiple imputation of unordered categorical data	112
5.1	Monotone missing data	112
5.2	Multivariate normal imputation for categorical data	114

5.3	Maximum indicant model	114
5.3.1	Continuous and categorical variable	117
5.3.2	Imputing missing data	119
5.3.3	More than one categorical variable	120
5.4	General location model	121
5.5	FCS with categorical data	122
5.6	Perfect prediction issues with categorical data	124
5.7	Software	126
5.8	Discussion	126
6	Nonlinear relationships	127
6.1	Passive imputation	128
6.2	No missing data in nonlinear relationships	130
6.3	Missing data in nonlinear relationships	133
6.3.1	Predictive Mean Matching (PMM)	133
6.3.2	Just Another Variable (JAV)	134
6.3.3	Joint modelling approach	135
6.3.4	Extension to more general models and missing data patterns	138
6.3.5	Metropolis-Hastings sampling	140
6.3.6	Rejection sampling	141
6.3.7	FCS approach	143
6.4	Discussion	145
7	Interactions	147
7.1	Interaction variables fully observed	147
7.2	Interactions of categorical variables	151
7.3	General nonlinear relationships	155
7.4	Software	163
7.5	Discussion	164
PART III	ADVANCED TOPICS	165
8	Survival data, skips and large datasets	167
8.1	Time-to-event data	167
8.1.1	Imputing missing covariate values	169
8.1.2	Survival data as categorical	173
8.1.3	Imputing censored survival times	177
8.2	Nonparametric, or 'hot deck' imputation	180
8.2.1	Nonparametric imputation for survival data	182
8.3	Multiple imputation for skips	184
8.4	Two-stage MI	188
8.5	Large datasets	190
8.5.1	Large datasets and joint modelling	190

8.5.2	Shrinkage by constraining parameters	192
8.5.3	Comparison of the two approaches	195
8.6	Multiple imputation and record linkage	195
8.7	Measurement error	197
8.8	Multiple imputation for aggregated scores	200
8.9	Discussion	202
9	Multilevel multiple imputation	203
9.1	Multilevel imputation model	203
9.2	MCMC algorithm for imputation model	214
9.3	Imputing level-2 covariates using FCS	220
9.4	Individual patient meta-analysis	222
9.4.1	When to apply Rubin's rules	224
9.5	Extensions	225
9.5.1	Random level-1 covariance matrices	226
9.5.2	Model fit	228
9.6	Discussion	228
10	Sensitivity analysis: MI unleashed	229
10.1	Review of MNAR modelling	230
10.2	Framing sensitivity analysis	233
10.3	Pattern mixture modelling with MI	235
10.3.1	Missing covariates	240
10.3.2	Application to survival analysis	241
10.4	Pattern mixture approach with longitudinal data via MI	246
10.4.1	Change in slope post-deviation	247
10.5	Piecing together post-deviation distributions from other trial arms	249
10.6	Approximating a selection model by importance weighting	257
10.6.1	Algorithm for approximate sensitivity analysis by re-weighting	259
10.7	Discussion	268
11	Including survey weights	269
11.1	Using model based predictions	269
11.2	Bias in the MI variance estimator	271
11.2.1	MI with weights	274
11.2.2	Estimation in domains	276
11.3	A multilevel approach	277
11.4	Further developments	280
11.5	Discussion	281
12	Robust multiple imputation	282
12.1	Introduction	282
12.2	Theoretical background	284

12.2.1	Simple estimating equations	284
12.2.2	The Probability Of Missingness (POM) model	285
12.2.3	Augmented inverse probability weighted estimating equation	286
12.3	Robust multiple imputation	287
12.3.1	Univariate MAR missing data	287
12.3.2	Longitudinal MAR missing data	289
12.4	Simulation studies	292
12.4.1	Univariate MAR missing data	292
12.4.2	Longitudinal monotone MAR missing data	293
12.4.3	Longitudinal nonmonotone MAR missing data	293
12.4.4	Nonlongitudinal nonmonotone MAR missing data	297
12.4.5	Results and discussion	297
12.5	The RECORD study	302
12.6	Discussion	304
Appendix A Markov Chain Monte Carlo		306
Appendix B Probability distributions		310
B.1	Posterior for the multivariate normal distribution	313
Bibliography		316
Index of Authors		327
Index of Examples		332
Index		334