



JAMES R. CARPENTER, JONATHAN W. BARTLETT,
TIM P. MORRIS, ANGELA M. WOOD,
MATTEO QUARTAGNO, MICHAEL G. KENWARD

Multiple Imputation and its Application

Second Edition

STATISTICS IN PRACTICE

WILEY

Table of Contents

[Cover](#)

[Title Page](#)

[Copyright](#)

[Preface to the second edition](#)

[Data acknowledgements](#)

[Acknowledgements](#)

[Glossary](#)

[Part I: Foundations](#)

[1 Introduction](#)

[1.1 Reasons for missing data](#)

[1.2 Examples](#)

[1.3 Patterns of missing data](#)

[1.4 Inferential framework and notation](#)

[1.5 Using observed data to inform assumptions
about the missingness mechanism](#)

[1.6 Implications of missing data mechanisms
for regression analyses](#)

[1.7 Summary](#)

[Exercises](#)

[2 The multiple imputation procedure and its
justification](#)

[2.1 Introduction](#)

[2.2 Intuitive outline of the MI procedure](#)

[2.3 The generic MI procedure](#)

[2.4 Bayesian justification of MI](#)

[2.5 Frequentist inference](#)

[2.6 Choosing the number of imputations](#)

[2.7 Some simple examples](#)

[2.8 MI in more general settings](#)

[2.9 Constructing congenial imputation models](#)

[2.10 Discussion](#)

[Exercises](#)

[Part II: Multiple Imputation for Simple Data Structures](#)

[3 Multiple imputation of quantitative data](#)

[3.1 Regression imputation with a monotone missingness pattern](#)

[3.2 Joint modelling](#)

[3.3 Full conditional specification](#)

[3.4 Full conditional specification versus joint modelling](#)

[3.5 Software for multivariate normal imputation](#)

[3.6 Discussion](#)

[Exercises](#)

[4 Multiple imputation of binary and ordinal data](#)

[4.1 Sequential imputation with monotone missingness pattern](#)

[4.2 Joint modelling with the multivariate normal distribution](#)

[4.3 Modelling binary data using latent normal variables](#)

[4.4 General location model](#)

[4.5 Full conditional specification](#)

[4.6 Issues with over-fitting](#)

[4.7 Pros and cons of the various approaches](#)

[4.8 Software](#)

[4.9 Discussion](#)

[Exercises](#)

[Note](#)

[5 Imputation of unordered categorical data](#)

[5.1 Monotone missing data](#)

[5.2 Multivariate normal imputation for categorical data](#)

[5.3 Maximum indicant model](#)

[5.4 General location model](#)

[5.5 FCS with categorical data](#)

[5.6 Perfect prediction issues with categorical data](#)

[5.7 Software](#)

[5.8 Discussion](#)

[Exercises](#)

[Part III: Multiple Imputation in Practice](#)

[6 Non-linear relationships, interactions, and other derived variables](#)

[6.1 Introduction](#)

[6.2 No missing data in derived variables](#)

[6.3 Simple methods](#)

[6.4 Substantive-model-compatible imputation](#)

[6.5 Returning to the problems](#)

[Exercises](#)

[7 Survival data](#)

[7.1 Missing covariates in time-to-event data](#)

[7.2 Imputing censored event times](#)

[7.3 Non-parametric, or 'hot deck' imputation](#)

[7.4 Case-cohort designs](#)

7.5 Discussion

Exercises

8 Prognostic models, missing data, and multiple imputation

8.1 Introduction

8.2 Motivating example

8.3 Missing data at model implementation

8.4 Multiple imputation for prognostic modelling

8.5 Model building

8.6 Model performance

8.7 Model validation

8.8 Incomplete data at implementation

Exercises

9 Multi-level multiple imputation

9.1 Multi-level imputation model

9.2 MCMC algorithm for imputation model

9.3 Extensions

9.4 Other imputation methods

9.5 Individual participant data meta-analysis

9.6 Software

9.7 Discussion

Exercises

10 Sensitivity analysis: MI unleashed

10.1 Review of MNAR modelling

10.2 Framing sensitivity analysis: estimands

10.3 Pattern mixture modelling with MI

10.4 Pattern mixture approach with longitudinal data via MI

[10.5 Reference based imputation](#)

[10.6 Approximating a selection model by importance weighting](#)

[10.7 Discussion](#)

[Exercises](#)

[11 Multiple imputation for measurement error and misclassification](#)

[11.1 Introduction](#)

[11.2 Multiple imputation with validation data](#)

[11.3 Multiple imputation with replication data](#)

[11.4 External information on the measurement process](#)

[11.5 Discussion](#)

[Exercises](#)

[12 Multiple imputation with weights](#)

[12.1 Using model-based predictions in strata](#)

[12.2 Bias in the MI variance estimator](#)

[12.3 MI with weights](#)

[12.4 A multi-level approach](#)

[12.5 Further topics](#)

[12.6 Discussion](#)

[Exercises](#)

[Note](#)

[13 Multiple imputation for causal inference](#)

[13.1 Multiple imputation for causal inference in point exposure studies](#)

[13.2 Multiple imputation and propensity scores](#)

[13.3 Principal stratification *via* multiple imputation](#)

[13.4 Multiple imputation for IV analysis](#)

[13.5 Discussion](#)

[Exercises](#)

[14 Using multiple imputation in practice](#)

[14.1 A general approach](#)

[14.2 Objections to multiple imputation](#)

[14.3 Reporting of analyses with incomplete data](#)

[14.4 Presenting incomplete baseline data](#)

[14.5 Model diagnostics](#)

[14.6 How many imputations?](#)

[14.7 Multiple imputation for each substantive model, project, or dataset?](#)

[14.8 Large datasets](#)

[14.9 Multiple imputation and record linkage](#)

[14.10 Setting random number seeds for multiple imputation analyses](#)

[14.11 Simulation studies including multiple imputation](#)

[14.12 Discussion](#)

[Exercises](#)

[Note](#)

[Appendix A Markov Chain Monte Carlo](#)

[A.1 Metropolis Hastings sampler](#)

[A.2 Gibbs sampler](#)

[A.3 Missing data](#)

[Appendix B Probability distributions](#)

[B.1 Posterior for the multivariate normal distribution](#)

[Appendix C Overview of multiple imputation in R, Stata](#)

[C.1 Basic multiple imputation using R](#)

[C.2 Basic MI using Stata](#)

[References](#)

[Author Index](#)

[Index of Examples](#)

[Subject Index](#)

[End User License Agreement](#)

List of Tables

Chapter 1

[Table 1.1 YCS variables for exploring the relationship between Year 11 att...](#)

[Table 1.2 Pattern of missing values in the YCS data.](#)

[Table 1.3 Asthma study: withdrawal pattern by treatment arm.](#)

[Table 1.4 Three variables: all possible missing value patterns.](#)

[Table 1.5 Asthma study: mean \$FEV_{1}\$ \(litres\) at each visit, by dropout pattern a...](#)

[Table 1.6 Coefficients \(standard errors\), and receiver operating characteri...](#)

[Table 1.7 Asthma study: estimated treatment effect fitting treatment, and t...](#)

[Table 1.8 Typical two-by-two table of counts relating outcome to exposure....](#)

[Table 1.9 Missing data mechanisms and bias on coefficient estimates with ty...](#)

[Table 1.10 Empirical illustration of Table 1.9 using logistic regression.](#)

[Table 1.11 Missing data patterns for sub-sample ignorable likelihood. As be...](#)

[Table 1.12 Results of fictional blood pressure survey. Each cell is the tal...](#)

Chapter 2

[Table 2.1 Simple example of MI; imputed values shown in bold.](#)

[Table 2.2 Description of variables in the datasets *ch2full.txt*, *ch2miss.txt*...](#)

Chapter 3

[Table 3.1 Coefficient estimates from regressions on observed data in placeb...](#)

[Table 3.2 Results of multiple imputation for the placebo arm of the asthma ...](#)

Chapter 4

[Table 4.1 Withdrawal pattern for dental data, for observations up to 6 hour...](#)

[Table 4.2 Results of multiple imputation for the estimation of the treatmen...](#)

Chapter 5

[Table 5.1 Missingness pattern in the 62,578 Youth Cohort Study \(YCS\) record...](#)

[Table 5.2 Complete records analysis of the Youth Cohort Study data and anal...](#)

Chapter 6

[Table 6.1 An example dataset with binary composite outcomes where \$Y_b\$ is MCAR...](#)

[Table 6.2 Asthma data: estimates of 12-week treatment effect on \$FEV_{1}\$ \(litres\),...](#)

[Table 6.3 Results of simulation study with quadratic effect of \$X\$ on \$Y\$.](#)

[Table 6.4 Baseline characteristics of \$n = 383\$ ADNI participants with mild cogniti...](#)

[Table 6.5 Estimated log-hazard ratios \(SEs\) for Cox proportional hazards mo...](#)

Chapter 7

[Table 7.1 Adjusted excess hazard ratio \(EHR\) of death in the first year fol...](#)

[Table 7.2 Cox proportional hazards outcome model simulation results.](#)

[Table 7.3 Results from applying MI methods to the ARIC cohort with case-coh...](#)

Chapter 9

[Table 9.1 Class size data: parameter estimates \(standard errors\) from fitti...](#)

[Table 9.2 Class size data: parameter estimates from original data, and with...](#)

[Table 9.3 Summary of data used for analysis of effectiveness of the PAR for...](#)

[Table 9.4 Results of fitting \(9.6\) and \(9.7\) to complete records and after ...](#)

Chapter 10

[Table 10.1 Review quality index of paper 1 by whether or not paper 2 was re...](#)

[Table 10.2 Peer review trial: inference for comparison of self-taught packa...](#)

[Table 10.3 Youth cohort study: model for the differences in GCSE points sco...](#)

[Table 10.4 Occupation of parents of students of Bangladeshi ethnicity, in c...](#)

[Table 10.5 Illustration of sensitivity analysis for increased mortality amo...](#)

[Table 10.6 Asthma study: estimates of the treatment effect in the hypotheti...](#)

[Table 10.7 Estimated 12-week treatment effect on \$FEV_{1}\$ \(litres\), from ANCOVA, m...](#)

[Table 10.8 Results of simulation study. Estimated standard errors of the es...](#)

[Table 10.9 Parameter estimates \(log odds ratios\) for logistic regression fo...](#)

[Table 10.10 Peer review trial: estimated effects of the self-training inter...](#)

[Table 10.11 Description of variables in the reviewer_sensitivity.dta datase...](#)

[Table 10.12 Summary results for the asthma study: the outcome is \$FEV_{1}\$ \(litres\)...](#)

Chapter 11

[Table 11.1 Results of the simulation study with misclassified binary exposu...](#)

[Table 11.2 Distribution of, and missing values in, baseline variables from ...](#)

[Table 11.3 Log hazard ratio estimates and 95% confidence intervals for the ...](#)

Chapter 12

[Table 12.1 Simulation results: mean, empirical SE, multiple imputation SE,...](#)

[Table 12.2 Analysis of millennium cohort study: parameter estimates and ass...](#)

Chapter 13

[Table 13.1 Data from a fictitious study with \$Z\$ indicating treatment, \$Y\$ obse...](#)

Appendix C

[Table C.1 Depiction of Stata's `mi` styles.](#)

List of Illustrations

Chapter 1

[Figure 1.1 Detail from a senior mandarin's house front in New Territories, H...](#)

[Figure 1.2 Plot of 200 hypothetical incomes against job type.](#)

[Figure 1.3 Contour-plot of the difference in average \(litres\) between activ...](#)

[Figure 1.4 Asthma study: mean \(litres\) at each visit, by randomised arm for...](#)

[Figure 1.5 Boxplot of GCSE points by whether parental occupation is observed...](#)

[Figure 1.6 Regression lines for simulated data. \(a\) Fitted regression line t...](#)

[Figure 1.7 Missing covariates: effect of different mechanisms. \(a\) Effect of...](#)

Chapter 2

[Figure 2.1 Relative efficiency \(\) from 2.28, for various values of and](#)

[Figure 2.2 Fill in your estimates of the mean of y1 and its 95% confidence i...](#)

Chapter 3

[Figure 3.1 Post burn-in trace plots for and corresponding elements of the c...](#)

Chapter 4

[Figure 4.1 Adaptive rounding threshold, as \$\bar{Y}\$ varies.](#)

[Figure 4.2 Plot of \$\Phi\(0.6x\)\$ versus \$\expit\(x\)\$ and line of...](#)

Chapter 6

[Figure 6.1 \(a\) A strong quadratic relationship with two values of missing \(...](#)

[Figure 6.2 Predictive mean matching under two missing-at-random mechanisms. ...](#)

[Figure 6.3 Plot of baseline \(litres\) versus 12 week \(litres\) by treatment ...](#)

[Figure 6.4 Some examples of 'FP2' fractional polynomial curves with powers ,...](#)

[Figure 6.5 Histogram of GCSE points score, for pupils who obtained one or mo...](#)

Chapter 7

[Figure 7.1 Probability of missingness for covariates as a function of age.](#)

Chapter 9

[Figure 9.1 Example of trace plot and autocorrelation plot from R package mit...](#)

Chapter 10

[Figure 10.1 Schematic illustration of increasing the rate of decline by pos...](#)

[Figure 10.2 Mean FE \(litres\) against time for the four different intercurrent...](#)

[Figure 10.3 Information about the sample mean varies with \(recall informati...](#)

[Figure 10.4 Simulation results for sensitivity analysis by re-weighting in l...](#)

Chapter 13

[Figure 13.1 Directed acyclic graph depicting as an instrumental variable fo...](#)

Chapter 14

[Figure 14.1 An UpSet plot for the fictional 'heart attack data'; details in ...](#)

[Figure 14.2 Illustration of Monte Carlo error in estimating the odds ratio a...](#)

[Figure 14.3 Schematic for imputation of longitudinal clinical record data us...](#)

Statistics in Practice

Founding Editor

Vic Barnett

Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title's special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceuticals; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

Multiple Imputation and its Application

Second Edition

James R. Carpenter

London School of Hygiene & Tropical Medicine, London, UK

MRC Clinical Trials Unit, University College London, UK

Jonathan W. Bartlett

London School of Hygiene & Tropical Medicine, London, UK

Tim P. Morris

MRC Clinical Trials Unit, University College London, UK

Angela M. Wood

University of Cambridge, Cambridge, UK

Matteo Quartagno

MRC Clinical Trials Unit at University College London, UK

Michael G. Kenward

Ashkirk, UK

WILEY

This edition first published 2023

© 2023 John Wiley & Sons Ltd

Edition History: John Wiley & Sons Ltd (1e, 2013)

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of James R. Carpenter, Jonathan W. Bartlett, Tim P. Morris, Angela M. Wood, Matteo Quartagno and Michael G. Kenward to be identified as the authors of this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher

nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data applied for

ISBN: 9781119756088 (Hardback); 9781119756095 (Adobe PDF);
9781119756101 (ePub)

Cover design: Wiley

Cover image: Courtesy of Harvey Goldstein

Preface to the second edition

No study of any complexity manages to collect all the intended data. Analysis of the resulting partially collected data must therefore address the issues raised by the missing data. Beyond simply estimating the proportion of missing values, the interplay between the substantive questions and the reasons for the missing data is crucial. There is no simple, universal, solution.

Suppose, for a substantive question at hand, the consequences of missing data in terms of bias and loss of precision are non-trivial. Then the analyst must make a set of assumptions about the reasons, or mechanisms, causing data to be missing, and perform an inferentially valid analysis under these assumptions. In this regard, analysis of a partially observed dataset is the same as any statistical analysis; the difference is that when data are missing we cannot assess the validity of these assumptions in the way we might do in a regression analysis, for example. Hence, sensitivity analysis, where we explore the robustness of inference to different assumptions about the reasons for missing data, is important.

Given a set of assumptions about the reasons data are missing, there are a number of statistical methods for carrying out the analysis. These include the expectation-maximization (EM) algorithm, inverse probability (of non-missingness) weighting, a full Bayesian analysis and, depending on the setting, a direct application of maximum likelihood. These methods, and those derived from them, each have their own advantages in particular settings. We focus on multiple imputation for its practical utility, broad applicability, and relatively straightforward application. Since the first edition was published ten years ago, new

applications of multiple imputation have continued to emerge and we have had to be selective in what we cover. The topics included are those we have found most relevant for our research and teaching.

Like the first edition, the book is divided into three parts. [Part I](#) lays the foundations, with an introductory chapter outlining the issues raised by missing data, followed by a chapter describing the theoretical foundations of multiple imputation. [Part II](#) describes the application of multiple imputation for standard regression analyses, explaining how MI can be used for continuous, categorical, and ordinal data. [Part III](#) describes how to apply MI in a range of practical settings, specifically analysis with non-linear relationships, analysis of survival data, development and validation of prognostic models, analysis with multilevel data structures, sensitivity analysis, handling measurement error, analysis involving weights, and causal inference. We conclude with a chapter outlining some broad practical points on the application of multiple imputation. While readers may wish to read only specific relevant chapters in [Part III](#), [Chapter 14](#) is intended to be relevant to all readers. We illustrate ideas with a range of examples from the medical and social sciences, reflecting the wide application that MI has seen in recent years.

Each chapter concludes with a range of exercises, designed to consolidate and deepen understanding of the material. The computer-based exercises have been designed with R and Stata users in mind. The book's home page at <https://missingdata.lshtm.ac.uk> contains both (i) hints for the exercises (including suggestions for R and Stata code) and (ii) full solutions where applicable.

We welcome feedback from readers. Please email james.carpenter@lshtm.ac.uk in the first instance.

*James R. Carpenter, Jonathan W. Bartlett, Tim P. Morris,
Angela M. Wood, Matteo Quartagno and Mike G.*

Kenward

September 2022

Data acknowledgements

We are grateful to the following:

1. AstraZeneca for permission to use data from the 5-arm asthma study in examples in [Chapters 1, 3, 6, and 10](#);
2. GlaxoSmithKline for permission to use data from the dental pain study in [Chapter 4](#);
3. Mike English (Director, Child and Newborn Health Group, Nairobi, Kenya) for permission to data from a multi-faceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals in [Chapter 9](#);
4. Peter Blatchford for permission to use data from the Class Size Study (Blatchford *et al.*, [2002](#)) in [Chapter 9](#);
5. Sara Schroter for permission to use data from the study to improve the quality of peer review in [Chapter 10](#).

In [Chapter 12](#), we used data from the Millennium Cohort Study available through the UK Data Service (ukdataservice.ac.uk), study number 2000031.

In [Chapters 1, 5, 9, and 10](#), we have analysed data from the Youth Cohort Time Series for England, Wales, and Scotland, 1984–2002 First Edition, Colchester, Essex, published by and freely available from the UK Data Archive, Study Number SN 5765. Thanks to Vernon Gayle for introducing us to these data.

In [Chapter 6](#), we have analysed data from the Alzheimer's Disease Neuro-imaging Initiative (ADNI) database (<https://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of

ADNI and/or provided data but did not participate in analysis or writing of this book. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health

Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

In [Chapter 11](#), we have analysed data from the US National Health and Nutrition Examination Survey (NHANES), which is freely available from the US Centers for Disease Control and Prevention. We thank the participants and survey organisers of NHANES.

Acknowledgements

No book of this kind is written in a vacuum, and we are grateful to many friends and colleagues.

In particular, we would like to thank the following for discussions and/or inputs that have contributed to this book or shaped our thinking through the years: John Carlin, Orlagh Carroll, Stephen Cole, Suzie Cro, Rhian Daniel, Dan Jackson, Ruth Keogh, Katherine Lee, Clemence Leyrat, Geert Molenberghs, Tra My Pham, Freya Roberts, James Roger, Jeremy Saxe, Shaun Seaman, Jonathan Sterne, Kate Tilling, and Ian White.

Specific thanks to Tetiana Gorbach for checking the exercises in [Chapter 2](#).

James would like to acknowledge many years' collaboration with Harvey Goldstein, who died in 2020.

James would like to thank Mike Elliott, Rod Little, Trivellore Raghunathan, and Jeremy Taylor for facilitating a visit to the Institute for Social Research and Department of Biostatistics at the University of Michigan, Ann Arbor, in Summer 2011, when the first draft of the first edition was written.

Thanks to Tim Collier for the anecdote in [Section 1.3](#).

We also gratefully acknowledge funding support from the ESRC (three-year fellowship for James Carpenter, RES-063-27-0257, and follow-on funding RES-189-25-0103) and MRC (grants G0900724, G0900701 and G0600599, MC_UU_12023/21, MC_UU_12023/29, MC_UU_00004/07, MC_UU_00004/09 and MC_UU_00004/06, MR/T023953/1).

We would also like to thank Richard Davies and Kathryn Sharples at Wiley for their encouragement and support

Sensitivity analysis, [10](#), [21](#), [128](#), [360](#)

change of slope, [264](#)-267

estimands, [249](#)-251

framing, [249](#)-251

hypothetical estimand, [266](#)-267, [271](#)-275

information anchoring, [275](#)-278

longitudinal data, [263](#)-275

missing covariates, [256](#)-258

motivation, [245](#)-246

pattern mixture, [251](#)-254

policy estimand, [271](#)-275

selection model via re-weighting, [279](#)-289

reliability of, [281](#)-284

survival data, [187](#), [260](#)-263

Sequential regression (multivariate) imputation, see [Full conditional specification \(FCS\)](#).

Shared parameter model, [246](#)

Skewed continuous variables, [115](#)

Skips, [169](#)-172

interactions, [170](#)

Rubin's variance formula conservative, [172](#)

variable fully observed, [171](#)

variable partially observed, [171](#)

vary size of imputed datasets, [172](#)

Software, [4](#)

REALCOM-impute, [93](#), [103](#), [116](#), [127](#), [130](#), [387](#)

CAT for categorical data, [130](#)

MIX for mix of data types, [130](#)

smcfcs (R, Stata), [161](#)

winBUGS, [287](#)

binary data, [116](#)

categorical data, [130](#)

general location model, [123](#)

latent normal model, [130](#)

multivariate normal data, [93](#)

R, [93](#), [394](#)

SAS, [93](#)

Stata, [93](#), [396](#)

Squares, [137](#)-138

Statistical analysis plan, [363](#)

Substantive model, [40](#), [41](#), [44](#), [45](#), [47](#), [50](#), [63](#), [67](#)

estimating equations, [317](#)

interaction & non-linear, [148](#), [149](#)

Substantive model compatible imputation, [152](#)-165

case-cohort data, [193](#), [195](#)-197

competing risks data, [186](#)

multilevel data, [222](#)

survey data, [325](#)

survival data, [181](#)

time-varying covariate effects, [186](#)

Sum scores, [139](#), [141](#)

Survey, [4](#), [5](#), [39](#), [71](#)

auxiliary variables, [319](#)

condition to eliminate bias in Rubin's variance formula, [316](#)

design and estimated weights, [318](#)

domains, [71](#), [328](#)

impute separately in strata, [313](#)-314

linear estimators, [314](#), [319](#)

logistic regression, [322](#)

missing covariates, [320](#)

missing values in weight model, [329](#)

model based predictions, [313](#)

parameters vary across strata, [313](#)

two stage MI, [329](#)

weights form strata for multilevel MI, [320](#)-322

Survival data, [72](#), [175](#), [176](#)
as categorical, [180](#), [181](#)
baseline hazard, [183](#)
censoring mechanisms, [175](#)
competing risks, [176](#)
cumulative hazard as covariate, [176](#)-181
grouped, [180](#)
imputing censored, [179](#), [186](#)-191
misclassified covariates, [297](#)
missing covariates, [176](#)-181
Nelson-Aalen estimator, [179](#), [180](#)
normalising transformation, [179](#)
simulation study, [178](#)-179
time-varying covariates, [179](#)

Survivor function, [187](#)

Tobit model, [179](#)

Transform then impute, [143](#)-145

Transformation
normalising, [49](#), [93](#), [179](#)

Unblinding, [264](#)

Unit, [3](#), [5](#)-6

Wald tests, [55](#)

Weibull distribution, [178](#)

Wishart distribution, [85](#), [102](#), [226](#), [232](#), [325](#)