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

List of Tables

List of Illustrations

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 pharmaceutics; 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

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

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.

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