JAMES R. CARPENTER
MICHAEL G. KENWARD

Multiple Imputation and its Application

WILEY

STATISTICS IN PRACTICE

Multiple Imputation and its Application

James R. Carpenter

and

Michael G. Kenward

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

WILEY

A John Wiley & Sons, Ltd., Publication

This edition first published 2013

Registered office John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Carpenter, James R.
Multiple imputation and its application / James R. Carpenter, Michael G. Kenward. – 1st ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-74052-1 (hardback)
I. Kenward, Michael G., 1956- II. Title.
[DNLM: 1. Data Interpretation, Statistical. 2. Biomedical Research–methods. WA 950]
610.72′4–dc23

2012028821

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-74052-1

Cover photograph courtesy of Harvey Goldstein

Set in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Singapore by Markono Print Media Pte Ltd

Multiple Imputation and its Application

Statistics in Practice

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 clear explanations 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 fields of employment and research environments. Subject areas covered include medicine and pharmaceuticals; industry, finance and commerce; public services, and the earth and environmental sciences.

The books also provide support to students studying applied statistics courses in these areas. The demand for applied statistics graduates in these areas has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written textbooks to meet everyday practical needs. Feedback from readers will be valuable in monitoring our success.

A complete list of titles in this series appears at the end of the volume.

Preface

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. Unfortunately, the inferential consequences of missing data are not simply restricted to the proportion of missing observations. Instead, the interplay between the substantive questions and the reasons for the missing data is crucial. Thus, there is no simple, universal, solution.

Suppose, for the substantive question at hand, the inferential consequences of missing data are nontrivial. 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 EM algorithm, inverse probability 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. Nevertheless, we argue that none shares the practical utility, broad applicability and relative simplicity of Rubin's Multiple Imputation (MI).

Following an introductory chapter outlining the issues raised by missing data, the focus of this book is therefore MI. We outline its theoretical basis, and then describe its application to a range of common analysis in the medical and social sciences, reflecting the wide application that MI has seen in recent years. In particular, we describe its application with nonlinear relationships and interactions, with survival data and with multilevel data. The last three chapters consider practical sensitivity analyses, combining MI with inverse probability weighting, and doubly robust MI.

Self-evidently, a key component of an MI analysis, is the construction of an appropriate method of imputation. There is no unique, ideal, way in which this should be done. In particular, there there has been some discussion in the literature about the relative merits of the joint modelling and full conditional

specification approaches. We have found that thinking in terms of joint models is both natural and convenient for formulating imputation models, a range of which can then be (approximately) implemented using a full conditional specification approach. Differences in computational speed between joint modelling and full conditional specification are generally due to coding efficiency, rather than intrinsic superiority of one method over the other.

Throughout the book we illustrate the ideas with several examples. The code used for these examples, in various software packages, is available from the book's home page, which is at `http://www.wiley.com/go/multiple_imputation`, together with exercises to go with each chapter.

We welcome feedback from readers; any comments and corrections should be e-mailed to mi@lshtm.ac.uk. Unfortunately, we cannot promise to respond individually to each message.

Data acknowledgements

We are grateful to the following:

AstraZeneca for permission to use data from the 5-arm asthma study in examples in Chapters 1, 3, 7 and 10;

GlaxoSmithKline for permission to use data from the dental pain study in Chapter 4, and the RECORD study in Chapter 12;

Mike English (Director, Child and Newborn Health Group, Kemri-Wellcome Trust Research Programme, Nairobi, Kenya) for permission to use data from a multifaceted intervention to implement guidelines and improve admission paediatric care in Kenyan district hospitals, in Chapter 9;

Peter Blatchford for permission to use data from the Class Size Study (Blatchford *et al*, 2002) in Chapter 9, and

Sarah Schroter for permission to use data from the study to improve the quality of peer review in Chapter 10.

In Chapters 1, 5, 8, 10 and 11 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. We thank 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 (adni.loni.ucla.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 the analysis or in the writing of this book. A complete listing of ADNI investigators can be found at: `http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and nonprofit organisations, as a $60 million, five-year public-private partnership. 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). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, with approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see `www.adni-info.org`.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec, Inc.; Bristol-Myers Squibb Company; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development, LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Servier; Synarc, Inc.; and Takeda Pharmaceutical Company. 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro-imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

In Chapter 7 we have analysed data from the 1958 National Childhood Development Study. This is published, and freely available from the UK Data Archive, Study Number SN 5565 (waves 0–3) and SN 5566 (wave 4). We thank Ian Plewis for introducing us to these data.

Acknowledgements

No book of this kind is written in a vacuum, and we are grateful to many friends and colleagues for research collaborations, stimulating discussions and comments on draft chapters.

In particular we would like to thank members of the Missing Data Imputation and Analysis (MiDIA) group, including (in alphabetical order) Jonathan Bartlett, John Carlin, Rhian Daniel, Dan Jackson, Shaun Seaman, Jonathan Sterne, Kate Tilling and Ian White.

We would also like to acknowledge many years of collaboration with Geert Molenberghs, James Roger and Harvey Goldstein.

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 majority of the first draft was written.

Thanks to Tim Collier for the anecdote in §1.3.

We also gratefully acknowledge funding support from the ESRC (3-year fellowship for James Carpenter, RES-063-27-0257, and follow-on funding RES-189-25-0103) and MRC (grants G0900724, G0900701 and G0600599).

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

Lastly, thanks to our families for their forbearance and understanding over the course of this project.

Despite the encouragement and support of those listed above, the text inevitably contains errors and shortcomings, for which we take full responsibility.

James Carpenter and Mike Kenward
London School of Hygiene & Tropical Medicine

Glossary

Indices and symbols

i	indexes units, often individuals, unless defined otherwise
j	indexes variables in the data set, unless defined otherwise
n	total number of units in the data set, unless defined otherwise
p	depending on context, number of variables in a data set or number of parameters in a statistical model
X, Y, Z	random variables
$Y_{i,j}$	i^{th} observation on j^{th} variable, $i = 1, \ldots, n,\ j = 1, \ldots, p$.
θ	generic parameter
$\boldsymbol{\theta}$	generic parameter column vector, typically $p \times 1$
β, γ, δ	regression coefficients
$\boldsymbol{\beta}$	column vector of regression coefficients, typically $p \times 1$.

Matrices

$\boldsymbol{\Omega}$	matrix, typically of dimension $p \times p$.
$\boldsymbol{\Omega}_{i,j}$	i, j^{th} element of $\boldsymbol{\Omega}$
$\boldsymbol{\Omega}^T$	transpose of $\boldsymbol{\Omega}$, so that $\boldsymbol{\Omega}^T_{i,j} = \boldsymbol{\Omega}_{j,i}$.
$\mathbf{Y}_j = (Y_{1,j}, \ldots, Y_{n,j})^T$	$n \times 1$ column vector of observations on variable j.
$\text{tr}(\boldsymbol{\Omega})$	sum of diagonal elements of $\boldsymbol{\Omega}$, *ie* $\sum \boldsymbol{\Omega}_{i,i}$ known as the trace of the matrix.

Abbreviations

AIPW	Augmented Inverse Probability Weighting
CAR	Censoring At Random
CNAR	Censoring Not At Random
EM	Expectation Maximisation
FCS	Full Conditional Specification
FEV_1	Forced Expiratory Volume in 1 second (measured in litres)
FMI	Fraction of Missing Information
IPW	Inverse Probability Weighting

MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Multiple Imputation
MNAR	Missing Not At Random
POD	Partially Observed Data
POM	Probability Of Missingness
S.E.	Standard error

Probability distributions

$f(.)$	probability distribution function
$F(.)$	cumulative distribution function
'|'	to be verbalised 'given', as in $f(Y\|X)$
	'the probability distribution function of Y given X'

Contents