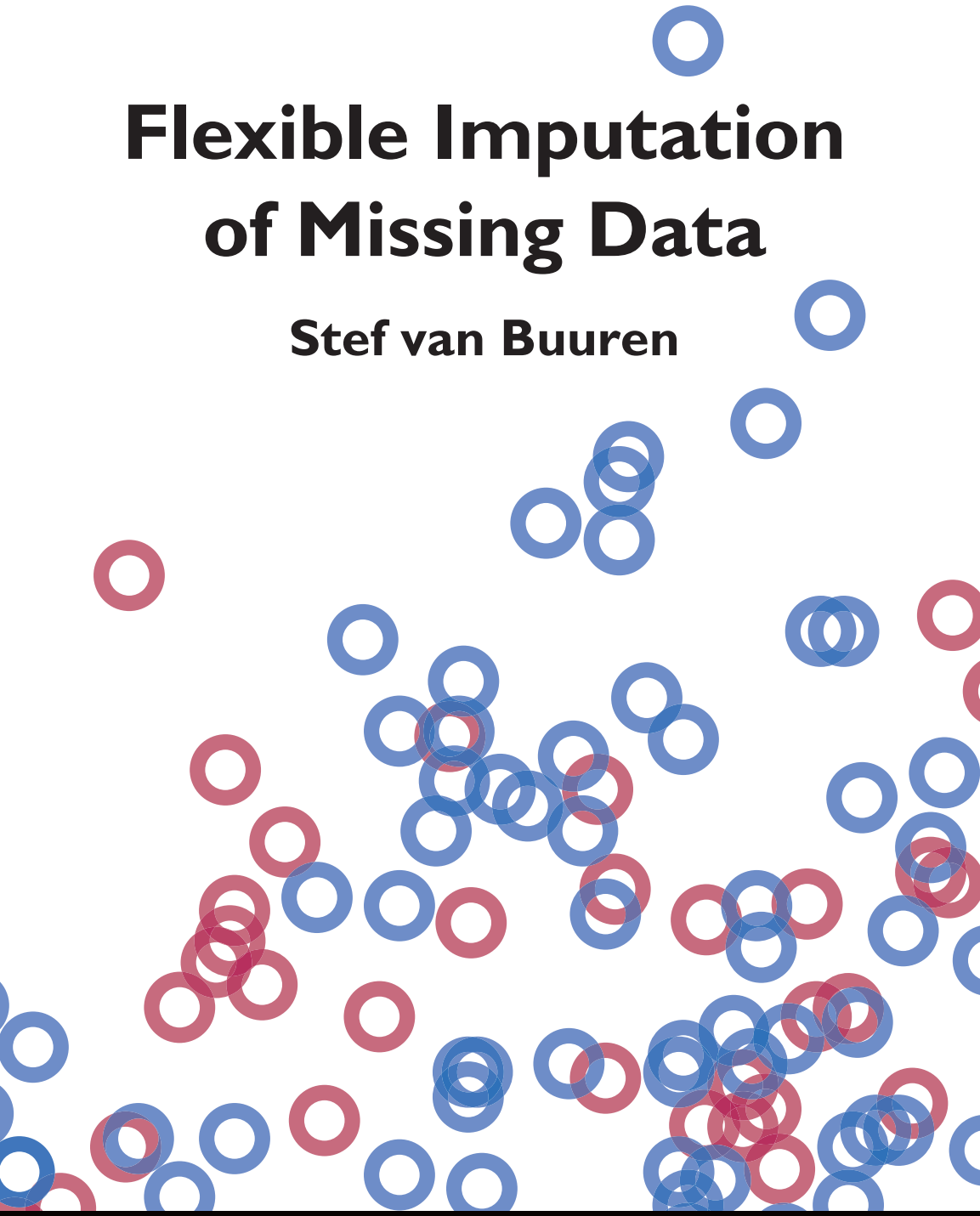
Chapman & Hall/CRC
Interdisciplinary Statistics Series

Flexible Imputation of Missing Data

Stef van Buuren



CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Flexible Imputation of Missing Data

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Published titles

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES	S. Lele and J. Richtsmeier
ASTROSTATISTICS	G. Babu and E. Feigelson
BAYESIAN ANALYSIS FOR POPULATION ECOLOGY	Ruth King, Byron J.T. Morgan, Olivier Gimenez, and Stephen P. Brooks
BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY	Andrew B. Lawson
BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY	S. Patterson and B. Jones
CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION	S. Green, J. Benedetti, and A. Smith
CLUSTER RANDOMISED TRIALS	R.J. Hayes and L.H. Moulton
CORRESPONDENCE ANALYSIS IN PRACTICE, SECOND EDITION	M. Greenacre
DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION	D.L. Fairclough
DYNAMICAL SEARCH	L. Pronzato, H.Wynn, and A. Zhigljavsky
FLEXIBLE IMPUTATION OF MISSING DATA	S. van Buuren
GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS	A. Skrondal and S. Rabe-Hesketh
GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA	K. Basford and J. Tukey
MARKOV CHAIN MONTE CARLO IN PRACTICE	W. Gilks, S. Richardson, and D. Spiegelhalter

Published titles

INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES	M. Waterman
MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS	P. Gustafson
MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS	J. P. Buonaccorsi
META-ANALYSIS OF BINARY DATA USING PROFILE LIKELIHOOD	D. Böhning, R. Kuhnert, and S. Rattanasiri
STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA	T. Speed
STATISTICAL AND COMPUTATIONAL PHARMACOGENOMICS	R. Wu and M. Lin
STATISTICS IN MUSICOLOGY	J. Beran
STATISTICS OF MEDICAL IMAGING	T. Lei
STATISTICAL CONCEPTS AND APPLICATIONS IN CLINICAL MEDICINE	J. Aitchison, J.W. Kay, and I.J. Lauder
STATISTICAL AND PROBABILISTIC METHODS IN ACTUARIAL SCIENCE	P.J. Boland
STATISTICAL DETECTION AND SURVEILLANCE OF GEOGRAPHIC CLUSTERS	P. Rogerson and I. Yamada
STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY	A. Bailer and W. Piegorsch
STATISTICS FOR FISSION TRACK ANALYSIS	R.F. Galbraith
VISUALIZING DATA PATTERNS WITH MICROMAPS	D.B. Carr and L.W. Pickle

This page intentionally left blank

Chapman & Hall/CRC
Interdisciplinary Statistics Series

Flexible Imputation of Missing Data

Stef van Buuren
TNO
Leiden, The Netherlands

University of Utrecht
The Netherlands

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Cover design by Celine Ostendorf.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20120227

International Standard Book Number-13: 978-1-4398-6825-6 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Voor Eveline, Guus, Otto en Maaike

This page intentionally left blank

Contents

Foreword xvii

Preface xix

About the Author xxi

Symbol Description xxiii

List of Algorithms xxv

I Basics 1

1 Introduction 3
1.1 The problem of missing data . . . 3
1.1.1 Current practice . . . 3
1.1.2 Changing perspective on missing data . . . 5
1.2 Concepts of MCAR, MAR and MNAR . . . 6
1.3 Simple solutions that do not (always) work . . . 8
1.3.1 Listwise deletion . . . 8
1.3.2 Pairwise deletion . . . 9
1.3.3 Mean imputation . . . 10
1.3.4 Regression imputation . . . 11
1.3.5 Stochastic regression imputation . . . 13
1.3.6 LOCF and BOFC . . . 14
1.3.7 Indicator method . . . 15
1.3.8 Summary . . . 15
1.4 Multiple imputation in a nutshell . . . 16
1.4.1 Procedure . . . 16
1.4.2 Reasons to use multiple imputation . . . 17
1.4.3 Example of multiple imputation . . . 18
1.5 Goal of the book . . . 20
1.6 What the book does not cover . . . 20
1.6.1 Prevention . . . 21
1.6.2 Weighting procedures . . . 21
1.6.3 Likelihood-based approaches . . . 22
1.7 Structure of the book . . . 23
1.8 Exercises . . . 23

2 Multiple imputation **25**
2.1 Historic overview . . . 25
2.1.1 Imputation . . . 25
2.1.2 Multiple imputation . . . 25
2.1.3 The expanding literature on multiple imputation . . . 27
2.2 Concepts in incomplete data . . . 28
2.2.1 Incomplete data perspective . . . 28
2.2.2 Causes of missing data . . . 29
2.2.3 Notation . . . 30
2.2.4 MCAR, MAR and MNAR again . . . 31
2.2.5 Ignorable and nonignorable ♠ . . . 33
2.2.6 Implications of ignorability . . . 34
2.3 Why and when multiple imputation works . . . 35
2.3.1 Goal of multiple imputation . . . 35
2.3.2 Three sources of variation ♠ . . . 36
2.3.3 Proper imputation . . . 38
2.3.4 Scope of the imputation model . . . 40
2.3.5 Variance ratios ♠ . . . 41
2.3.6 Degrees of freedom ♠ . . . 42
2.3.7 Numerical example . . . 43
2.4 Statistical intervals and tests . . . 44
2.4.1 Scalar or multi-parameter inference? . . . 44
2.4.2 Scalar inference . . . 44
2.5 Evaluation criteria . . . 45
2.5.1 Imputation is not prediction . . . 45
2.5.2 Simulation designs and performance measures . . . 47
2.6 When to use multiple imputation . . . 48
2.7 How many imputations? . . . 49
2.8 Exercises . . . 51

3 Univariate missing data **53**
3.1 How to generate multiple imputations . . . 53
3.1.1 Predict method . . . 55
3.1.2 Predict + noise method . . . 55
3.1.3 Predict + noise + parameter uncertainty . . . 55
3.1.4 A second predictor . . . 56
3.1.5 Drawing from the observed data . . . 56
3.1.6 Conclusion . . . 56
3.2 Imputation under the normal linear normal . . . 57
3.2.1 Overview . . . 57
3.2.2 Algorithms ♠ . . . 57
3.2.3 Performance . . . 59
3.2.4 Generating MAR missing data . . . 63
3.2.5 Conclusion . . . 64
3.3 Imputation under non-normal distributions . . . 65

3.3.1 Overview . . . 65
3.3.2 Imputation from the t-distribution ♠ . . . 66
3.3.3 Example ♠ . . . 67
3.4 Predictive mean matching . . . 68
3.4.1 Overview . . . 68
3.4.2 Computational details ♠ . . . 70
3.4.3 Algorithm ♠ . . . 73
3.4.4 Conclusion . . . 74
3.5 Categorical data . . . 75
3.5.1 Overview . . . 75
3.5.2 Perfect prediction ♠ . . . 76
3.6 Other data types . . . 78
3.6.1 Count data . . . 78
3.6.2 Semi-continuous data . . . 79
3.6.3 Censored, truncated and rounded data . . . 79
3.7 Classification and regression trees . . . 82
3.7.1 Overview . . . 82
3.7.2 Imputation using CART models . . . 83
3.8 Multilevel data . . . 84
3.8.1 Overview . . . 84
3.8.2 Two formulations of the linear multilevel model ♠ . . . 85
3.8.3 Computation ♠ . . . 86
3.8.4 Conclusion . . . 87
3.9 Nonignorable missing data . . . 88
3.9.1 Overview . . . 88
3.9.2 Selection model . . . 89
3.9.3 Pattern-mixture model . . . 90
3.9.4 Converting selection and pattern-mixture models . . . 90
3.9.5 Sensitivity analysis . . . 92
3.9.6 Role of sensitivity analysis . . . 93
3.10 Exercises . . . 93

4 Multivariate missing data **95**
4.1 Missing data pattern . . . 95
4.1.1 Overview . . . 95
4.1.2 Summary statistics . . . 96
4.1.3 Influx and outflux . . . 99
4.2 Issues in multivariate imputation . . . 101
4.3 Monotone data imputation . . . 102
4.3.1 Overview . . . 102
4.3.2 Algorithm . . . 103
4.4 Joint modeling . . . 105
4.4.1 Overview . . . 105
4.4.2 Continuous data ♠ . . . 105
4.4.3 Categorical data . . . 107

4.5 Fully conditional specification . . . 108
4.5.1 Overview . . . 108
4.5.2 The MICE algorithm . . . 109
4.5.3 Performance . . . 111
4.5.4 Compatibility ♠ . . . 111
4.5.5 Number of iterations . . . 112
4.5.6 Example of slow convergence . . . 113
4.6 FCS and JM . . . 116
4.6.1 Relations between FCS and JM . . . 116
4.6.2 Comparison . . . 117
4.6.3 Illustration . . . 117
4.7 Conclusion . . . 121
4.8 Exercises . . . 121

5 Imputation in practice **123**
5.1 Overview of modeling choices . . . 123
5.2 Ignorable or nonignorable? . . . 125
5.3 Model form and predictors . . . 126
5.3.1 Model form . . . 126
5.3.2 Predictors . . . 127
5.4 Derived variables . . . 129
5.4.1 Ratio of two variables . . . 129
5.4.2 Sum scores . . . 132
5.4.3 Interaction terms . . . 133
5.4.4 Conditional imputation . . . 133
5.4.5 Compositional data ♠ . . . 136
5.4.6 Quadratic relations ♠ . . . 139
5.5 Algorithmic options . . . 140
5.5.1 Visit sequence . . . 140
5.5.2 Convergence . . . 142
5.6 Diagnostics . . . 146
5.6.1 Model fit versus distributional discrepancy . . . 146
5.6.2 Diagnostic graphs . . . 146
5.7 Conclusion . . . 151
5.8 Exercises . . . 152

6 Analysis of imputed data **153**
6.1 What to do with the imputed data? . . . 153
6.1.1 Averaging and stacking the data . . . 153
6.1.2 Repeated analyses . . . 154
6.2 Parameter pooling . . . 155
6.2.1 Scalar inference of normal quantities . . . 155
6.2.2 Scalar inference of non-normal quantities . . . 155
6.3 Statistical tests for multiple imputation . . . 156
6.3.1 Wald test ♠ . . . 157

6.3.2 Likelihood ratio test ♠ . . . 157
6.3.3 χ^2-test ♠ . . . 159
6.3.4 Custom hypothesis tests of model parameters ♠ . . . 159
6.3.5 Computation . . . 160
6.4 Stepwise model selection . . . 162
6.4.1 Variable selection techniques . . . 162
6.4.2 Computation . . . 163
6.4.3 Model optimism . . . 164
6.5 Conclusion . . . 166
6.6 Exercises . . . 166

II Case studies 169

7 Measurement issues 171
7.1 Too many columns . . . 171
7.1.1 Scientific question . . . 172
7.1.2 Leiden 85+ Cohort . . . 172
7.1.3 Data exploration . . . 173
7.1.4 Outflux . . . 175
7.1.5 Logged events . . . 176
7.1.6 Quick predictor selection for wide data . . . 177
7.1.7 Generating the imputations . . . 179
7.1.8 A further improvement: Survival as predictor variable 180
7.1.9 Some guidance . . . 181
7.2 Sensitivity analysis . . . 182
7.2.1 Causes and consequences of missing data . . . 182
7.2.2 Scenarios . . . 184
7.2.3 Generating imputations under the δ-adjustment . . . 185
7.2.4 Complete data analysis . . . 186
7.2.5 Conclusion . . . 187
7.3 Correct prevalence estimates from self-reported data . . . 188
7.3.1 Description of the problem . . . 188
7.3.2 Don't count on predictions . . . 189
7.3.3 The main idea . . . 190
7.3.4 Data . . . 191
7.3.5 Application . . . 192
7.3.6 Conclusion . . . 193
7.4 Enhancing comparability . . . 194
7.4.1 Description of the problem . . . 194
7.4.2 Full dependence: Simple equating . . . 195
7.4.3 Independence: Imputation without a bridge study . . . 196
7.4.4 Fully dependent or independent? . . . 198
7.4.5 Imputation using a bridge study . . . 199
7.4.6 Interpretation . . . 202
7.4.7 Conclusion . . . 203

7.5 Exercises 204

8 Selection issues **205**
8.1 Correcting for selective drop-out 205
8.1.1 POPS study: 19 years follow-up 205
8.1.2 Characterization of the drop-out 206
8.1.3 Imputation model 207
8.1.4 A degenerate solution 208
8.1.5 A better solution 210
8.1.6 Results 211
8.1.7 Conclusion 211
8.2 Correcting for nonresponse 212
8.2.1 Fifth Dutch Growth Study 212
8.2.2 Nonresponse 213
8.2.3 Comparison to known population totals 213
8.2.4 Augmenting the sample 214
8.2.5 Imputation model 215
8.2.6 Influence of nonresponse on final height 217
8.2.7 Discussion 218
8.3 Exercises 219

9 Longitudinal data **221**
9.1 Long and wide format 221
9.2 SE Fireworks Disaster Study 223
9.2.1 Intention to treat 224
9.2.2 Imputation model 225
9.2.3 Inspecting imputations 227
9.2.4 Complete data analysis 228
9.2.5 Results from the complete data analysis 229
9.3 Time raster imputation 230
9.3.1 Change score 231
9.3.2 Scientific question: Critical periods 232
9.3.3 Broken stick model ♠ 234
9.3.4 Terneuzen Birth Cohort 236
9.3.5 Shrinkage and the change score ♠ 237
9.3.6 Imputation 238
9.3.7 Complete data analysis 240
9.4 Conclusion 242
9.5 Exercises 244

III Extensions **247**

10 Conclusion **249**
10.1 Some dangers, some do's and some don'ts 249
10.1.1 Some dangers 249
10.1.2 Some do's 250
10.1.3 Some don'ts 251
10.2 Reporting 251
10.2.1 Reporting guidelines 252
10.2.2 Template 254
10.3 Other applications 255
10.3.1 Synthetic datasets for data protection 255
10.3.2 Imputation of potential outcomes 255
10.3.3 Analysis of coarsened data 256
10.3.4 File matching of multiple datasets 256
10.3.5 Planned missing data for efficient designs 256
10.3.6 Adjusting for verification bias 257
10.3.7 Correcting for measurement error 257
10.4 Future developments 257
10.4.1 Derived variables 257
10.4.2 Convergence of MICE algorithm 257
10.4.3 Algorithms for blocks and batches 258
10.4.4 Parallel computation 258
10.4.5 Nested imputation 258
10.4.6 Machine learning for imputation 259
10.4.7 Incorporating expert knowledge 259
10.4.8 Distribution-free pooling rules 259
10.4.9 Improved diagnostic techniques 260
10.4.10 Building block in modular statistics 260
10.5 Exercises 260

A Software **263**
A.1 R 263
A.2 S-PLUS 265
A.3 Stata 265
A.4 SAS 266
A.5 SPSS 266
A.6 Other software 266

References **269**

This page intentionally left blank

Foreword

I'm delighted to see this new book on multiple imputation by Stef van Buuren for several reasons. First, to me at least, having another book devoted to multiple imputation marks the maturing of the topic after an admittedly somewhat shaky initiation. Stef is certainly correct when he states in Section 2.1.2: "The idea to create multiple versions must have seemed outrageous at that time [late 1970s]. Drawing imputations from a distribution, instead of estimating the 'best' value, was a severe breach with everything that had been done before." I remember how this idea of multiple imputation was even ridiculed by some more traditional statisticians, sometimes for just being "silly" and sometimes for being hopelessly inefficient with respect to storage demands and outrageously expensive with respect to computational requirements.

Some others of us foresaw what was happening to both (a) computational storage (I just acquired a 64 GB flash drive the size of a small finger for under $60, whereas only a couple of decades ago I paid over $2500 for a 120 KB hard-drive larger than a shoe box weighing about 10 kilos), and (b) computational speed and flexibility. To develop statistical methods for the future while being bound by computational limitations of the past was clearly inapposite. Multiple imputation's early survival was clearly due to the insight of a younger generation of statisticians, including many colleagues and former students, who realized future possibilities.

A second reason for my delight at the publication of this book is more personal and concerns the maturing of the author, Stef van Buuren. As he mentions, we first met through Jan van Rijckevorsel at TNO. Stef was a young and enthusiastic researcher there, who knew little about the kind of statistics that I felt was essential for making progress on the topic of dealing with missing data. But consider the progress over the decades starting with his earlier work on MICE! Stef has matured into an independent researcher making important and original contributions to the continued development of multiple imputation.

This book represents a "no nonsense" straightforward approach to the application of multiple imputation. I particularly like Stef's use of graphical displays, which are badly needed in practice to supplement the more theoretical discussions of the general validity of multiple imputation methods. As I have said elsewhere, and as implied by much of what is written by Stef, "It's not that multiple imputation is so good; it's really that other methods for addressing missing data are so bad." It's great to have Stef's book on mul-

tiple imputation, and I look forward to seeing more editions as this rapidly developing methodology continues to become even more effective at handling missing data problems in practice.

Finally, I would like to say that this book reinforces the pride of an academic father who has watched one of his children grow and develop. This book is a step in the growing list of contributions that Stef has made, and, I am confident, will continue to make, in methodology, computational approaches, and application of multiple imputation.

Donald B. Rubin

Preface

We are surrounded by missing data. Problems created by missing data in statistical analysis have long been swept under the carpet. These times are now slowly coming to an end. The array of techniques for dealing with missing data has expanded considerably during the last decades. This book is about one such method: multiple imputation.

Multiple imputation is one of the great ideas in statistical science. The technique is simple, elegant and powerful. It is simple because it fills the holes in the data with plausible values. It is elegant because the uncertainty about the unknown data is coded in the data itself. And it is powerful because it can solve "other" problems that are actually missing data problems in disguise.

Over the last 20 years, I have applied multiple imputation in a wide variety of projects. I believe the time is ripe for multiple imputation to enter mainstream statistics. Computers and software are now potent enough to do the required calculations with little effort. What is still missing is a book that explains the basic ideas and that shows how these ideas can be put into practice. My hope is that this book can fill this gap.

The text assumes familiarity with basic statistical concepts and multivariate methods. The book is intended for two audiences:

- (Bio)statisticians, epidemiologists and methodologists in the social and health sciences
- Substantive researchers who do not call themselves statisticians, but who possess the necessary skills to understand the principles and to follow the recipes

In writing this text, I have tried to avoid mathematical and technical details as much as possible. Formulas are accompanied by a verbal statement that explains the formula in layperson terms. I hope that readers less concerned with the theoretical underpinnings will be able to pick up the general idea. The more technical material is marked by a club sign ♠, and can be skipped on first reading.

I used various parts of the book to teach a graduate course on imputation techniques at the University of Utrecht. The basics are in Chapters 1–4. Lecturing this material takes about 10 hours. The lectures were interspersed with sessions in which the students worked out the exercises from the book.

This book owes much to the ideas of Donald Rubin, the originator of multiple imputation. I had the privilege of being able to talk, meet and work with

him on many occasions. His clear vision and deceptively simple ideas have been a tremendous source of inspiration. I am also indebted to Jan van Rijckevorsel for bringing me into contact with Donald Rubin, and for establishing the scientific climate at TNO in which our work on missing data techniques could prosper.

Many people have helped realize this project. I thank Nico van Meeteren and Michael Holewijn of TNO for their trust and support. I thank Peter van der Heijden of Utrecht University for his support. I thank Rob Calver and the staff at Chapman & Hall/CRC for their help and advice. Many colleagues have commented on part or all of the manuscript: Hendriek Boshuizen, Elise Dusseldorp, Karin Groothuis-Oudshoorn, Michael Hermanussen, Martijn Heymans, Nicholas Horton, Shahab Jolani, Gerko Vink, Ian White and the research master students of the Spring 2011 class. Their comments have been very valuable for detecting and eliminating quite a few glitches. I happily take the blame for the remaining errors and vagaries.

The major part of the manuscript was written during a six-month sabbatical leave. I spent four months in Krukö, Sweden, a small village of just eight houses. I thank Frank van den Nieuwenhuijzen and Ynske de Koning for making their wonderful green house available to me. It was the perfect tranquil environment that, apart from snowplowing, provided a minimum of distractions. I also spent two months at the residence of Michael Hermanussen and Beate Lohse-Hermanussen in Altenhof, Germany. I thank them for their hospitality, creativity and wit. It was a wonderful time.

Finally, I thank my family, in particular my beloved wife Eveline, for their warm and ongoing support, and for allowing me to devote time, often nights and weekends, to work on this book. Eveline liked to tease me by telling people that I was writing "a book that no one understands." I fear that her statement is accurate, at least for 99% of the people. My hope is that you, my dear reader, will belong to the remaining 1%.

Stef van Buuren

About the Author

Stef van Buuren is a statistician at the Netherlands Organization of Applied Scientific Research TNO, and professor of Applied Statistics in Prevention at Utrecht University. He is the originator of the MICE algorithm for multiple imputation of multivariate data, and he co-developed the `mice` package in `R`. More information can be found at `www.stefvanbuuren.nl`.

This page intentionally left blank

Symbol Description

Y	$n \times p$ matrix of partially observed sample data (2.2.3)
R	$n \times p$ matrix, 0–1 response indicator of Y (2.2.3)
X	$n \times q$ matrix of predictors, used for various purposes (2.2.3)
Y_{obs}	observed sample data, values of Y with $R = 1$ (2.2.3)
Y_{mis}	unobserved sample data, values of Y with $R = 0$ (2.2.3)
n	sample size (2.2.3)
m	number of multiple imputations (2.2.3)
ψ	parameters of the missing data model that relates Y to R (2.2.4)
θ	parameters of the scientifically interesting model for the full data Y (2.2.5)
Q	$k \times 1$ vector with k scientific estimands (2.3.1)
$\hat{Q}$	$k \times 1$ vector, estimator of Q calculated from a hypothetically complete sample (2.3.1)
U	$k \times k$ matrix, within-imputation variance due to sampling (2.3.1)
ℓ	imputation number, $\ell = 1, \ldots, m$ (2.3.2)
Y_ℓ	ℓth imputed dataset, $\ell = 1, \ldots, m$ (2.3.2)
$\bar{Q}$	$k \times 1$ vector, estimator of Q calculated from the incompletely observed sample (2.3.2)
$\bar{U}$	$k \times k$ matrix, estimator of U from the incomplete data (2.3.2)
B	$k \times k$ matrix, between-imputation variance due to nonresponse (2.3.2)
T	total variance of $(Q - \bar{Q})$, $k \times k$ matrix (2.3.2)
λ	proportion of the variance attributable to the missing data for a scalar parameter (2.3.5)
γ	fraction of information missing due to nonresponse (2.3.5)
r	relative increase of variance due to nonresponse for a scalar parameter (2.3.5)
$\bar{\lambda}$	λ for multivariate Q (2.3.5)
$\bar{r}$	r for multivariate Q (2.3.5)
ν_{old}	old degrees of freedom (2.3.6)
ν	adjusted degrees of freedom (2.3.6)
y	univariate Y (3.2.1)
y_{obs}	vector with n_1 observed data values in y (3.2.1)
y_{mis}	vector n_0 missing data values in y (3.2.1)
$\dot{y}$	vector n_0 imputed values in y (3.2.1)
X_{obs}	subset of n_1 rows of X for which y is observed (3.2.1)
X_{mis}	subset of n_0 rows of X for which y is missing (3.2.1)
$\hat{\beta}$	estimate of regression weight β (3.2)
$\dot{\beta}$	simulated regression weight for β (3.2.1)
$\hat{\sigma}^2$	estimate of residual variance σ^2 (3.2.1)

$\dot{\sigma}^2$ simulated residual variance for σ^2 (3.2.1)
κ ridge parameter (3.2.2)
η distance parameter in predictive mean matching (3.4.2)
$\hat{y}_i$ vector of n_1 predicted values given X_{obs} (3.4.2)
$\hat{y}_j$ vector of n_0 predicted values given X_{mis} (3.4.2)
C number of classes (3.8.2)
c class index, $c = 1, \ldots, C$ (3.8.2)
δ shift parameter in nonignorable models (3.9.1)
Y_j jth column of Y (4.1.1)
Y_{-j} all columns of Y except Y_j (4.1.1)
I_{jk} proportion of usable cases for imputing Y_j from Y_k (4.1.2)
O_{jk} proportion of observed cases in Y_j to impute Y_k (4.1.2)
I_j influx statistic to impute Y_j from Y_{-j} (4.1.3)
O_j outflux statistic to impute Y_{-j} from Y_j (4.1.3)
ϕ parameters of the imputation model that models the distribution of Y (4.3.1)
r_w r for Wald test (6.3.1)
D_w test statistic of Wald test (6.3.1)
ν_w ν for Wald test (6.3.1)
r_l r for likelihood ratio test (6.3.2)
D_l test statistic for likelihood ratio test (6.3.2)
ν_l ν for likelihood ratio test (6.3.2)
r_x r for χ^2-test (6.3.3)
ν_x ν for χ^2-test (6.3.3)
D_x test statistic for χ^2-test (6.3.3)

List of Algorithms

3.1 Bayesian imputation under the normal linear model.♠ 58
3.2 Imputation under the normal linear model with bootstrap.♠ . 59
3.3 Predictive mean matching with a Bayesian β and a stochastic matching distance (Type 1 matching).♠ 73
3.4 Imputation of a binary variable by means of Bayesian logistic regression.♠ 76
3.5 Imputation of right-censored data using predictive mean matching, Kaplan–Meier estimation and the bootstrap.♠ 82
3.6 Imputation under a tree model using the bootstrap.♠ 84
3.7 Multilevel imputation with bootstrap under the linear normal model with class-wise error variances.♠ 87
4.1 Monotone data imputation of multivariate missing data.♠ . . 104
4.2 Imputation of missing data by a joint model for multivariate normal data.♠ 106
4.3 MICE algorithm for imputation of multivariate missing data.♠ 110
5.1 Multiple imputation of quadratic terms.♠ 141

This page intentionally left blank

Part I

Basics

This page intentionally left blank

Chapter 1

Introduction

1.1 The problem of missing data

1.1.1 Current practice

The mean of the numbers 1, 2 and 4 can be calculated in R as

```
> y <- c(1, 2, 4)
> mean(y)
[1] 2.3
```

where y is a vector containing three numbers, and where mean(y) is the R expression that returns their mean. Now suppose that the last number is missing. R indicates this by the symbol NA, which stands for "not available":

```
> y <- c(1, 2, NA)
> mean(y)
[1] NA
```

The mean is now undefined, and R informs us about this outcome by setting the mean to NA. It is possible to add an extra argument na.rm = TRUE to the function call. This removes any missing data before calculating the mean:

```
> mean(y, na.rm = TRUE)
[1] 1.5
```

This makes it possible to calculate a result, but of course the set of observations on which the calculations are based has changed. This may cause problems in statistical inference and interpretation.

Similar problems occur in multivariate analysis. Many users of R will have seen the following error message:

```
> lm(Ozone ~ Wind, data = airquality)
Error in na.fail.default(list(Ozone = c(41, 36, 12, 18, NA,
  missing values in object
```

This code calls function `lm()` to fit a linear regression to predict daily ozone concentration (ppb) from wind speed (mph) using the built-in dataset `airquality`. The program cannot continue because the value of `Ozone` is unknown for some days. It is easy to omit any incomplete records by specifying the `na.action = na.omit` argument to `lm()`. The regression weights can now be obtained as

```
> fit <- lm(Ozone ~ Wind, data = airquality,
      na.action = na.omit)
> coef(fit)
(Intercept)        Wind
       96.9        -5.6
```

The `R` object `fit` stores the results of the regression analysis, and the `coef()` function extract the regression weights from it. In practice, it is cumbersome to supply the `na.action()` function each time. We can change the factory-fresh setting of the `options` as

```
> options(na.action = na.omit)
```

This command eliminates the error message once and for all. Users of other software packages like `SPSS`, `SAS` and `Stata` enjoy the "luxury" that this deletion option has already been set for them, so the calculations can progress silently. The procedure is known as *listwise deletion* or *complete case analysis*, and is widely used.

Though listwise deletion allows the calculations to proceed, it may cause problems in interpretation. For example, we can find the number of deleted cases in the fitted model as

```
> deleted <- na.action(fit)
> naprint(deleted)
[1] "37 observations deleted due to missingness"
```

The `na.action()` function finds the cases that are deleted from the fitted model. The `naprint()` function echoes the number of deleted cases. Now suppose we fit a better predictive model by including solar radiation (`Solar.R`) in the model. We obtain

```
> fit2 <- lm(Ozone ~ Wind + Solar.R, data = airquality)
> naprint(na.action(fit2))
[1] "42 observations deleted due to missingness"
```

where the previous three separate statements have been combined into one line. The number of deleted days increased from 37 to 42 since some of the days had no value for `Solar.R`. Thus, changing the model altered the sample.

There are methodological and statistical issues associated with this procedure. Some questions that come to mind are:

- Can we compare the regression coefficients from both models?
- Should we attribute differences in the coefficients to changes in the model or to changes in the subsample?
- Do the estimated coefficients generalize to the study population?
- Do we have enough cases to detect the effect of interest?
- Are we making the best use of the costly collected data?

Getting the software to run is one thing, but this alone does not address the questions caused by the missing data. This book discusses techniques that allow us to consider the type of questions raised above.

1.1.2 Changing perspective on missing data

The standard approach to missing data is to delete them. It is illustrative to search for missing values in published data. Hand et al. (1994) published a highly useful collection of small datasets across the statistical literature. The collection covers an impressive variety of topics. Only 13 out of the 510 datasets in the collection actually had a code for the missing data. In many cases, the missing data problem has probably been "solved" in some way, usually without telling us how many missing values there were originally. It is impossible to track down the original data for most datasets in Hand's book. However, we can easily do this for dataset number 357, a list of scores of 34 athletes in 10 sport events at the 1988 Olympic decathlon in Seoul. The table itself is complete, but a quick search on the Internet revealed that initially 39 instead of 34 athletes participated. Five of them did not finish for various reasons, including the dramatic disqualification of the German favorite Jürgen Hingsen because of three false starts in the 100-meter sprint. It is probably fair to assume that deletion occurred silently in many of the other datasets.

The inclination to delete the missing data is understandable. Apart from the technical difficulties imposed by the missing data, the occurrence of missing data has long been considered a sign of sloppy research. It is all too easy for a referee to write:

> This study is weak because of the large amount of missing data.

Publication chances are likely to improve if there is no hint of missingness. Orchard and Woodbury (1972, p. 697) remarked:

> Obviously the best way to treat missing data is not to have them.

Though there is a lot of truth in this statement, Orchard and Woodbury realized the impossibility of attaining this ideal in practice.

The prevailing scientific practice is to downplay the missing data. Reviews

on reporting practices are available in various fields: clinical trials (Wood et al., 2004), cancer research (Burton and Altman, 2004), educational research (Peugh and Enders, 2004), epidemiology (Klebanoff and Cole, 2008), developmental psychology (Jeliĉić et al., 2009), general medicine (Mackinnon, 2010) and developmental pediatrics (Aylward et al., 2010). The picture that emerges from these studies is quite consistent:

- The presence of missing data is often not explicitly stated in the text;
- Default methods like listwise deletion are used without mentioning them;
- Different tables are based on different sample sizes;
- Model-based missing data methods, such as direct likelihood, full information maximum likelihood and multiple imputation, are notably underutilized.

Missing data are there, whether we like it or not. In the social sciences, it is nearly inevitable that some respondents will refuse to participate or to answer certain questions. In medical studies, attrition of patients is very common. Allison (2002, p. 1) begins by observing:

> Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data.

Even the most carefully designed and executed studies produce missing values. The really interesting question is how we deal with incomplete data.

The theory, methodology and software for handling incomplete data problems have been vastly expanded and refined over the last decades. The major statistical analysis packages now have facilities for performing the appropriate analyses. This book aims to contribute to a better understanding of the issues involved, and provides a methodology for dealing with incomplete data problems in practice.

1.2 Concepts of MCAR, MAR and MNAR

Before we review a number of simple fixes for the missing data in Section 1.3 let us take a short look at the terms MCAR, MAR and MNAR. A more detailed definition of these concepts will be given later in Section 2.2.3. Rubin (1976) classified missing data problems into three categories. In his theory every data point has some likelihood of being missing. The process that governs these probabilities is called the *missing data mechanism* or *response mechanism*. The model for the process is called the *missing data model* or *response model*.

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing, apart from the obvious loss of information. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. Another example is when we take a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR. While convenient, MCAR is often unrealistic for the data at hand.

If the probability of being missing is the same only within groups defined by the *observed* data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and if we can assume MCAR *within* the type of surface, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability to be included depends on some known property. MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption.

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Rubin's distinction is important for understanding why some methods will not work. His theory lays down the conditions under which a missing data method can provide valid statistical inferences. Most simple fixes only work under the restrictive and often unrealistic MCAR assumption. If MCAR is implausible, such methods can provide biased estimates.

1.3 Simple solutions that do not (always) work

1.3.1 Listwise deletion

Complete case analysis (listwise deletion) is the default way of handling incomplete data in many statistical packages, including `SPSS`, `SAS` and `Stata`. The function `na.omit()` does the same in `S-PLUS` and `R`. The procedure eliminates all cases with one or more missing values on the analysis variables.

The major advantage of complete case analysis is convenience. If the data are MCAR, listwise deletion produces unbiased estimates of means, variances and regression weights. Under MCAR, listwise deletion produces standard errors and significance levels that are correct for the reduced subset of data, but that are often larger relative to all available data.

A disadvantage of listwise deletion is that it is potentially wasteful. It is not uncommon in real life applications that more than half of the original sample is lost, especially if the number of variables is large. King et al. (2001) estimated that the percentage of incomplete records in the political sciences exceeded 50% on average, with some studies having over 90% incomplete records. It will be clear that a smaller subsample could seriously degrade the ability to detect the effects of interest.

If the data are not MCAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations. Little and Rubin (2002, pp. 41–44) showed that the bias in the estimated mean increases with the difference between means of the observed and missing cases, and with the proportion of the missing data. Schafer and Graham (2002) reported an elegant simulation study that demonstrates the bias of listwise deletion under MAR and MNAR. However, complete case analysis is not always bad. The implications of the missing data are different depending on where they occur (outcomes or predictors), and the parameter and model form of the complete data analysis. In the context of regression analysis, listwise deletion possesses some unique properties that make it attractive in particular settings. There are cases in which listwise deletion can provide better estimates than even the most sophisticated procedures. Since their discussion requires a bit more background than can be given here, we defer the treatment to Section 2.6.

Listwise deletion can introduce inconsistencies in reporting. Since listwise deletion is automatically applied to the active set of variables, different analyses on the same data are often based on different subsamples. In principle, it is possible to produce one global subsample using all active variables. In practice, this is unattractive since the global subsample will always have fewer cases than each of the local subsamples, so it is common to create different subsets for different tables. It will be evident that this complicates their comparison and generalization to the study population.

In some cases, listwise deletion can lead to nonsensical subsamples. For example, the rows in the `airquality` dataset used in Section 1.1.1 correspond

to 154 consecutive days between May 1, 1973 and September 30, 1973. Deleting days affects the time basis. It would be much harder, if not impossible, to perform analyses that involve time, e.g., to identify weekly patterns or to fit autoregressive models that predict from previous days.

The opinions on the value of listwise deletion vary. Miettinen (1985, p. 231) described listwise deletion as

> ... the only approach that assures that no bias is introduced under any circumstances...

a bold statement, but incorrect. At the other end of the spectrum we find Enders (2010, p. 39):

> In most situations, the disadvantages of listwise deletion far outweigh its advantages.

Schafer and Graham (2002, p. 156) cover the middle ground:

> If a missing data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective.

The leading authors in the field are, however, wary of providing advice about the percentage of missing cases below which it is still acceptable to do listwise deletion. Little and Rubin (2002) argue that it is difficult to formulate rules of thumb since the consequences of using listwise deletion depend on more than the missing data rate alone. Vach (1994, p. 113) expressed his dislike for simplistic rules as follows:

> It is often supposed that there exists something like a critical missing rate up to which missing values are not too dangerous. The belief in such a global missing rate is rather stupid.

1.3.2 Pairwise deletion

Pairwise deletion, also known as *available-case analysis*, attempts to remedy the data loss problem of listwise deletion. The method calculates the means and (co)variances on all observed data. Thus, the mean of variable X is based on all cases with observed data on X, the mean of variable Y uses all cases with observed Y-values, and so on. For the correlation and covariance, all data are taken on which both X and Y have non-missing scores. Subsequently, the matrix of summary statistics are fed into a program for regression analysis, factor analysis or other modeling procedures.

We can calculate the mean, covariances and correlations of the `airquality` data under pairwise deletion in `R` as:

```
> mean(airquality, na.rm = TRUE)
> cor(airquality, use = "pair")
> cov(airquality, use = "pair")
```

SPSS, SAS and Stata contain many procedures with an option for pairwise deletion. The method is simple, uses all available information and produces consistent estimates of mean, correlations and covariances under MCAR (Little and Rubin, 2002, p. 55). Nevertheless, when taken together these estimates have major shortcomings. The estimates can be biased if the data are not MCAR. Furthermore, there are computational problems. The correlation matrix may not be positive definite, which is requirement for most multivariate procedures. Correlations outside the range [−1,+1] can occur, a problem that comes from different subsets used for the covariances and the variances. Such problems are more severe for highly correlated variables (Little, 1992). Another problem is that it is not clear which sample size should be used for calculating standard errors. Taking the average sample size yields standard errors that are too small (Little, 1992).

The idea behind pairwise deletion is to use all available information. Though this idea is good, the proper analysis of the pairwise matrix requires sophisticated optimization techniques and special formulas to calculate the standard errors (Van Praag et al., 1985; Marsh, 1998). Pairwise deletion should only be used if the procedure that follows it is specifically designed to take deletion into account. The attractive simplicity of pairwise deletion as a general missing data method is thereby lost.

1.3.3 Mean imputation

A quick fix for the missing data is to replace them by the mean. We may use the mode for categorical data. Suppose we want to impute the mean in Ozone and Solar.R of the airquality data. SPSS, SAS and Stata have prebuilt functions that substitute the mean. This book uses the R package mice (Van Buuren and Groothuis-Oudshoorn, 2011). This software is a contributed package that extends the functionality of R. Before mice can be used, it must be installed. An easy way to do this is to type:

```
> install.packages("mice")
```

which searches the Comprehensive R Archive Network (CRAN), and installs the requested package on the local computer. After successful installation, the mice package can be loaded by

```
> library("mice")
```

Imputing the mean in each variable can now be done by

```
> imp <- mice(airquality, method = "mean", m = 1,
      maxit = 1)
 iter imp variable
  1   1  Ozone  Solar.R
```

The argument method="mean" specifies mean imputation, the argument m=1 requests a single imputed dataset, and maxit=1 sets the number of iterations

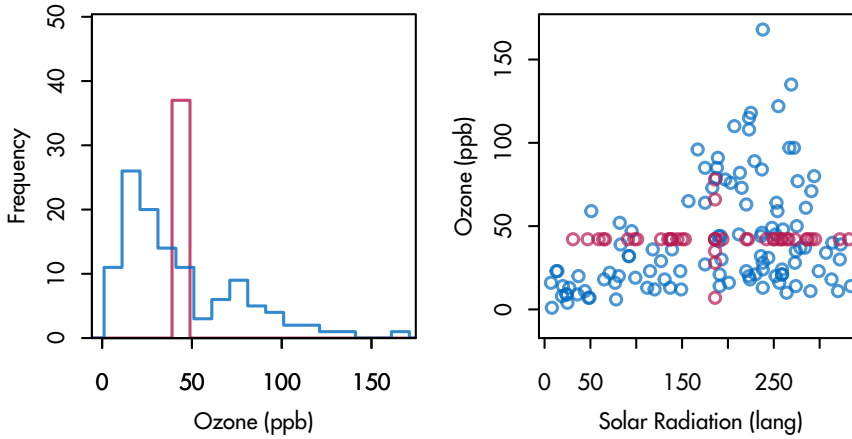


Figure 1.1: Mean imputation of `Ozone`. Blue indicates the observed data, red indicates the imputed values.

to 1 (no iteration). The latter two options options can be left to their defaults with essentially the same result.

Mean imputation distorts the distribution in several ways. Figure 1.1 displays the distribution of `Ozone` after imputation. The `mice` package adopts the Abayomi convention for the colors (Abayomi et al., 2008). Blue refers to the observed part of the data, red to the synthetic part of the data (also called the *imputed values* or *imputations*), and black to the combined data (also called the *imputed data* or *completed data*). The printed version of this book replaces blue by gray. In the figure on the left, the red bar at the mean stands out. Imputing the mean here actually creates a bimodal distribution. The standard deviation in the imputed data is equal to 28.7, much smaller than from the observed data alone, which is 33. The figure on the right-hand side shows that the relation between `Ozone` and `Solar.R` is distorted because of the imputations. The correlation drops from 0.35 in the blue points to 0.3 in the combined data.

Mean imputation is a fast and simple fix for the missing data. However, it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Mean imputation should perhaps only be used as a rapid fix when a handful of values are missing, and it should be avoided in general.

1.3.4 Regression imputation

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. The first step involves building a model

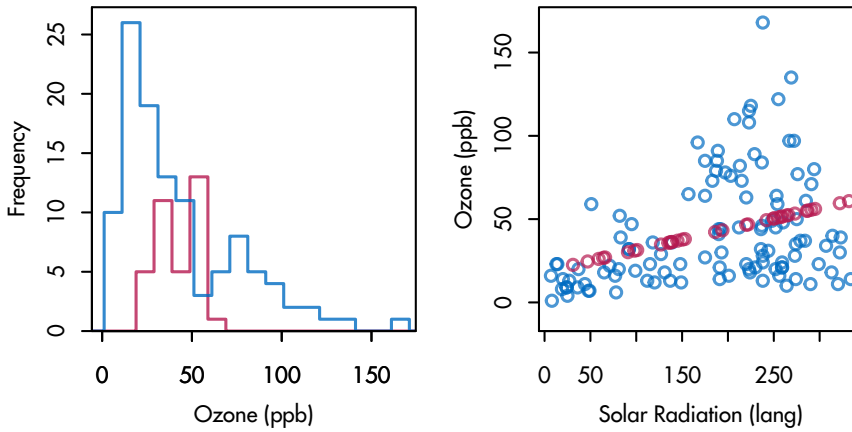


Figure 1.2: Regression imputation: Imputing `Ozone` from the regression line.

from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data. Suppose that we model `Ozone` by the linear regression function of `Solar.R`.

```
> fit <- lm(Ozone ~ Solar.R, data = airquality)
> pred <- predict(fit, newdata = ic(airquality))
```

Figure 1.2 shows the result. The imputed values correspond to the most likely values under the model. However, the ensemble of imputed values vary less than the observed values. It may be that each of the individual points is the best under the model, but it is very unlikely that the real (but unobserved) values of `Ozone` would have had this distribution. Imputing predicted values also has an effect on the correlation. The red points have a correlation of 1 since they are located on a line. If the red and blue dots are combined, then the correlation increases from 0.35 to 0.39.

Regression imputation yields unbiased estimates of the means under MCAR, just like mean imputation, and of the regression weights of the imputation model if the explanatory variables are complete. Moreover, the regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model. In the example this corresponds to the situation where `Solar.R` would explain any differences in the probability that `Ozone` is missing. On the other hand, the variability of the imputed data is systematically underestimated. The degree of underestimation depends on the explained variance and on the proportion of missing cases (Little and Rubin, 2002, p. 64).

Imputing predicted values can yield realistic imputations if the prediction is close to perfection. If so, the method reconstructs the missing parts from the available data. In essence, there was not really any information missing

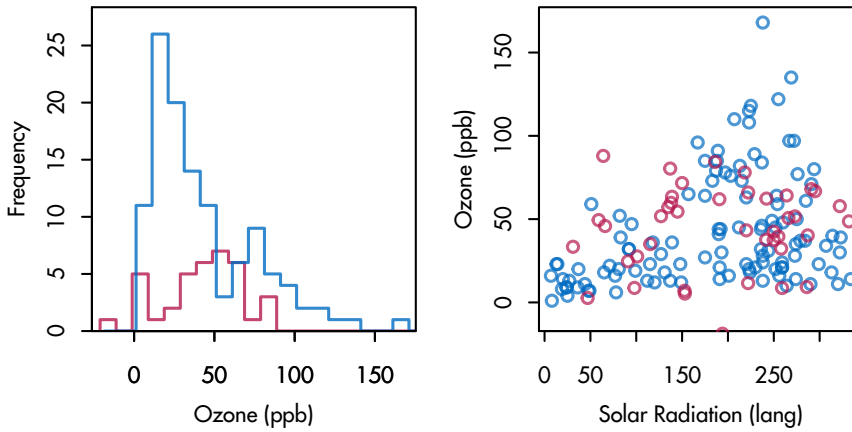


Figure 1.3: Stochastic regression imputation of `Ozone`.

in the first place, it was only coded in a different form. This type of missing data is unlikely to surface in most applications.

1.3.5 Stochastic regression imputation

Stochastic regression imputation is a refinement of regression imputation that adds noise to the predictions. This will have a downward effect on the correlation. We can impute `Ozone` by stochastic regression imputation as:

```
> imp <- mice(airquality[, 1:2], method = "norm.nob",
      m = 1, maxit = 1, seed = 1)
 iter imp variable
  1   1  Ozone  Solar.R
```

The `method="norm.nob"` argument requests a plain, non-Bayesian, stochastic regression method. This method first estimates the intercept, slope and residual variance under the linear model, then generates imputed value according to these specification. We will come back to the details in Section 3.2. The `seed` argument makes the solution reproducible. Figure 1.3 shows the results. The addition of noise to the predictions opens up the distribution of the imputed values, as intended.

Note that some new complexities arise. There are several imputations with negative values. Such values are implausible since negative `Ozone` concentrations do not exist in the real world. Also, the high end of the distribution is not well covered. The cause of this is that the relation in the observed data is somewhat heteroscedastic. The variability of `Ozone` seems to increase up to the solar radiation level of 250 langleys, and decreases after that. Though it is

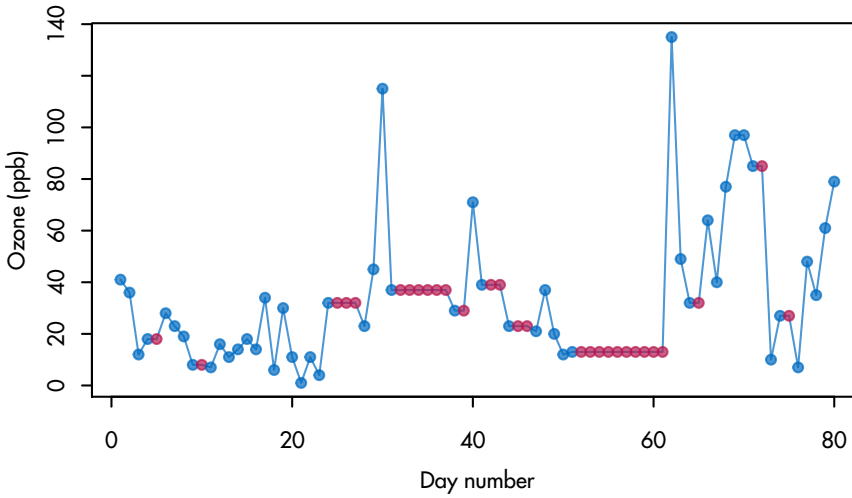


Figure 1.4: Imputation of `Ozone` by last observation carried forward (LOCF).

unclear whether this is a genuine meteorological phenomenon, the imputation model did not account for this feature.

Stochastic regression imputation is an important step forward. In particular it preserves not only the regression weights, but also the correlation between variables (cf. Exercise 3). Stochastic regression imputation does not solve all problems, and there are many subtleties that need to be addressed. However, the main idea to draw from the residuals is very powerful, and forms the basis of more advanced imputation techniques.

1.3.6 LOCF and BOFC

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) require longitudinal data. The idea is to take the last observed value as a replacement for the missing data. Figure 1.4 illustrates the method applied to the first 80 days of the `Ozone` series. The stretches of red dots indicate the imputations.

LOCF is convenient because it generates a complete dataset. The method is used in clinical trials. The U.S. Food and Drug Administration (FDA) has traditionally viewed LOCF as the preferred method of analysis, considering it conservative and less prone to selection than listwise deletion. However, Molenberghs and Kenward (2007, pp. 47–50) show that the bias can operate in both directions, and that LOCF can yield biased estimates even under MCAR. LOCF needs to be followed by a proper statistical analysis method that distinguishes between the real and imputed data. This is typically not done, however. Additional concerns about a reversal of the time direction are given in Kenward and Molenberghs (2009).

The Panel on Handling Missing Data in Clinical Trials recommends that LOCF and BOCF should not be used as the primary approach for handling missing data unless the assumptions that underlie them are scientifically justified (National Research Council, 2010, p. 77).

1.3.7 Indicator method

Suppose that we want to fit a regression, but there are missing values in one of the explanatory variables. The indicator method (Miettinen, 1985, p. 232) replaces each missing value by a zero and extends the regression model by the response indicator. The procedure is applied to each incomplete variable. The user analyzes the extended model instead of the original.

This method is popular in public health and epidemiology. An advantage is that the indicator method retains the full dataset. Also, it allows for systematic differences between the observed and the unobserved data by inclusion of the response indicator. However, the method can yield severely biased regression estimates, even under MCAR and for low amounts of missing data (Vach and Blettner, 1991; Greenland and Finkle, 1995; Knol et al., 2010).

On the other hand, White and Thompson (2005) point out that the method can be useful to estimate the treatment effect in randomized trials when a baseline covariate is partially observed. If the missing data are restricted to the covariate, if the interest is solely restricted to estimation of the treatment effect, if compliance to the allocated treatment is perfect and if the model is linear without interactions, then using the indicator method for that covariate yields an unbiased estimate of the treatment effect. This is true even if the missingness depends on the covariate itself.

The conditions under which the indicator method works are often difficult to achieve in practice. The method does not allow for missing data in the outcomes, both of which frequently occur in real data. While the indicator method may be suitable in some special cases, it falls short as a general way to treat missing data.

1.3.8 Summary

Table 1.1 provides a summary of the methods discussed in this section. The table addresses two topics: whether the method yields the correct results on average (unbiasedness), and whether it produces the correct standard error. Unbiasedness is evaluated with respect to the mean, the regression weight (of the regression with the incomplete variable as dependent) and the correlation.

The table identifies the assumptions on the missing data mechanism each method must make in order to produce unbiased estimates. Both deletion methods always require MCAR. In addition, for listwise deletion there are two MNAR special cases (cf. Section 2.6). Regression imputation and stochastic regression imputation can yield unbiased estimates under MAR. In order to

Table 1.1: Overview of assumptions made by simple methods

	Mean	Unbiased Reg Weight	Correlation	Standard Error
Listwise deletion	MCAR	MCAR	MCAR	Too large
Pairwise deletion	MCAR	MCAR	MCAR	Complicated
Mean imputation	MCAR	–	–	Too small
Regression imp	MAR	MAR	–	Too small
Stochastic imp	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

work, the model needs to be correctly specified. LOCF and the indicator method are incapable of providing consistent estimates, even under MCAR.

Listwise deletion produces standard errors that are correct for the subset of complete cases, but in general too large for the entire dataset. Calculation of standard errors under pairwise deletion is complicated. The standard errors after imputation are too small since the standard calculations make no distinction between the observed data and the imputed data. Correction factors for some situations have been developed (Schafer and Schenker, 2000), but a more convenient solution is multiple imputation.

1.4 Multiple imputation in a nutshell

1.4.1 Procedure

Multiple imputation creates $m > 1$ complete datasets. Each of these datasets is analyzed by standard analysis software. The m results are pooled into a final point estimate plus standard error by simple pooling rules ("Rubin's rules"). Figure 1.5 illustrates the three main steps in multiple imputation: imputation, analysis and pooling.

The analysis starts with observed, incomplete data. Multiple imputation creates several complete versions of the data by replacing the missing values by plausible data values. These plausible values are drawn from a distribution specifically modeled for each missing entry. Figure 1.5 portrays $m = 3$ imputed datasets. In practice, m is often taken larger (cf. Section 2.7). The number $m = 3$ is taken here just to make the point that the technique creates multiple versions of the imputed data. The three imputed datasets are identical for the observed data entries, but differ in the imputed values. The magnitude of these difference reflects our uncertainty about what value to impute.

The second step is to estimate the parameters of interest from each imputed dataset. This is typically done by applying the analytic method that we would

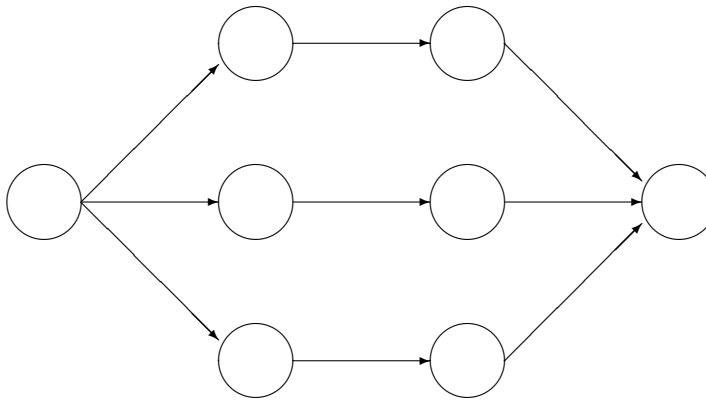


Figure 1.5: Scheme of main steps in multiple imputation.

have used had the data been complete. The results will differ because their input data differ. It is important to realize that these differences are caused only because of the uncertainty about what value to impute.

The last step is to pool the m parameter estimates into one estimate, and to estimate its variance. The variance combines the conventional sampling variance (within-imputation variance) and the extra variance caused by the missing data extra variance caused by the missing data (between-imputation variance). Under the appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties.

1.4.2 Reasons to use multiple imputation

Multiple imputation (Rubin, 1987a, 1996) solves the problem of "too small" standard errors in Table 1.1. Multiple imputation is unique in the sense that it provides a mechanism for dealing with the inherent uncertainty of the imputations themselves.

Our level of confidence in a particular imputed value is expressed as the variation across the m completed datasets. For example, in a disability survey, suppose that the respondent answered the item whether he could walk, but did not provide an answer to the item whether he could get up from a chair. If the person can walk, then it is highly likely that the person will also be able to get up from the chair. Thus, for persons who can walk, we can draw a "yes" for missing "getting up from a chair" with a high probability, say 0.99, and use the drawn value as the imputed value. In the extreme, if we are really certain, we always impute the same value for that person. More generally, we

are less confident about the true value. Suppose that, in a growth study, height is missing for a subject. If we only know that this person is a woman, this provides some information about likely values, but not so much. So the range of plausible values from which we draw is much larger here. The imputations for this woman will thus vary a lot over the different datasets. Multiple imputation is able to deal with both high-confidence and low-confidence situations equally well.

Another reason to use multiple imputation is that it separates the solution of the missing data problem from the solution of the complete data problem. The missing data problem is solved first, the complete data problem next. Though these phases are not completely independent, the answer to the scientifically interesting question is not obscured anymore by the missing data. The ability to separate the two phases simplifies statistical modeling, and hence contributes to a better insight into the phenomenon of scientific study.

1.4.3 Example of multiple imputation

Continuing with the `airquality` dataset, it is straightforward to apply multiple imputation. The following code imputes the missing data five times, fits a linear regression model to predict `Ozone` in each of the imputed datasets, and pools the five sets of estimated parameters.

```
> imp <- mice(airquality, seed = 1, print = FALSE)
> fit <- with(imp, lm(Ozone ~ Wind + Temp + Solar.R))
> tab <- round(summary(pool(fit)), 3)
> tab[, c(1:3, 5)]
               est     se    t Pr(>|t|)
(Intercept) -64.31 24.614 -2.6    0.015
Wind         -3.11  0.828 -3.8    0.003
Temp          1.64  0.243  6.8    0.000
Solar.R       0.05  0.023  2.1    0.037
```

Fitting the same model to the complete cases can be done by:

```
> fit <- lm(Ozone ~ Wind + Temp + Solar.R, data = airquality,
      na.action = na.omit)
> round(coef(summary(fit)), 3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -64.34     23.055    -2.8    0.006
Wind           -3.33      0.654    -5.1    0.000
Temp            1.65      0.254     6.5    0.000
Solar.R         0.06      0.023     2.6    0.011
```

The solutions are nearly identical here, which is due to the fact that most missing values occur in the outcome variable. The standard errors of the multiple imputation solution are slightly smaller than in the complete case analysis.

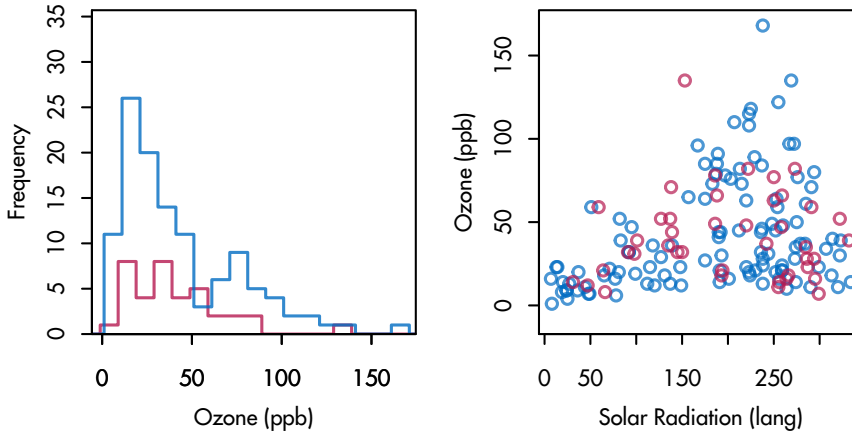


Figure 1.6: Multiple imputation of `Ozone`. Plotted are the imputed values from the first imputation.

It is often the case that multiple imputation is more efficient than complete case analysis. Depending on the data and the model at hand, the differences can be dramatic.

Figure 1.6 shows the distribution and scattergram for the observed and imputed data combined. The imputations are taken from the first completed dataset. The blue and red distributions are quite similar. Problems with the negative values as in Figure 1.3 are now gone since the imputation method used observed data as donors to fill the missing data. Section 3.4 describes the method in detail. Note that the red points respect the heteroscedastic nature of the relation between `Ozone` and `Solar.R`. All in all, the red points look as if they could have been measured if they had not been missing. The reader can easily recalculate the solution and inspect these plots for the other imputations.

Figure 1.7 plots the completed `Ozone` data. The imputed data of all five imputations are plotted for the days with missing `Ozone` scores. In order to avoid clutter, the lines that connect the dots are not drawn for the imputed values. Note that the pattern of imputed values varies substantially over the days. At the beginning of the series, the values are low and the spread is small, in particular for the cold and windy days 25–27. The small spread for days 25–27 indicates that the model is quite sure of these values. High imputed values are found around the hot and sunny days 35–42, whereas the imputations during the moderate days 52–61 are consistently in the moderate range. Note how the available information helps determine sensible imputed values that respect the relations between wind, temperature, sunshine and ozone.

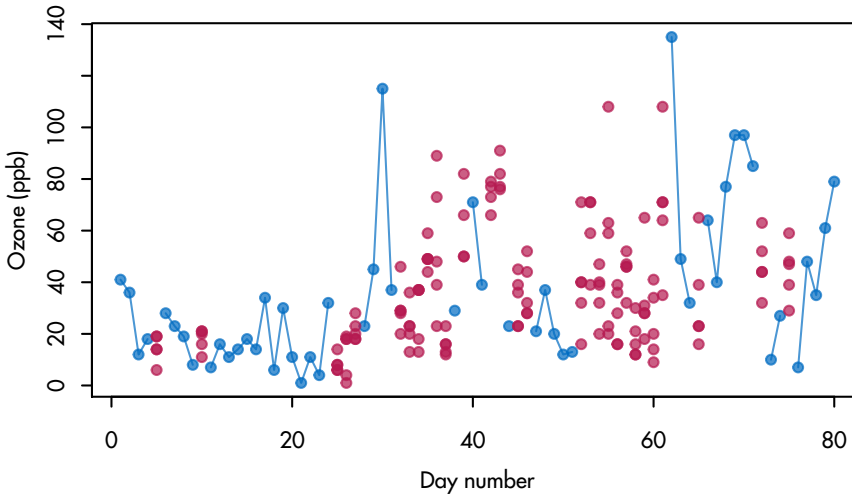


Figure 1.7: Multiple imputation of `Ozone`. Plotted are the observed values (in blue) and the multiply imputed values (in red). One red dot at (61,168) is not plotted.

1.5 Goal of the book

The main goal of this book is to add multiple imputation to the tool chest of practitioners. The text explains the ideas underlying multiple imputation, discusses when multiple imputation is useful, how to do it in practice and how to report the results of the steps taken.

The computations are done with the help of the `R` package `mice`, written by Karin Groothuis-Oudshoorn and myself (Van Buuren and Groothuis-Oudshoorn, 2011). The book thus also serves as an extended tutorial on the practical application of `mice`. Online materials that accompany the book can be found on www.multiple-imputation.com. My hope is that this hands-on approach will facilitate understanding of the key ideas in multiple imputation.

1.6 What the book does not cover

The field of missing data research is vast. This book focuses on multiple imputation. The book does not attempt cover the enormous body of literature on alternative approaches to incomplete data. This section briefly reviews three of these approaches.

1.6.1 Prevention

With the exception of McKnight et al. (2007, Chapter 4), books on missing data do not mention prevention. Yet, prevention of the missing data is the most direct attack on problems caused by the missing data. Prevention is fully in spirit with the quote of Orchard and Woodbury given on p. 5. There is a lot one could do to prevent missing data. The remainder of this section lists point-wise advice.

Minimize the use of intrusive measures, like blood samples. Visit the subject at home. Use incentives to stimulate response, and try to match up the interviewer and respondent on age and ethnicity. Adapt the mode of the study (telephone, face to face, web questionnaire, and so on) to the study population. Use a multi-mode design for different groups in your study. Quickly follow-up for people that do not respond, and where possible try to retrieve any missing data from other sources.

In experimental studies, try to minimize the treatment burden and intensity where possible. Prepare a well-thought-out flyer that explains the purpose and usefulness of your study. Try to organize data collection through an authority, e.g., the patient's own doctor. Conduct a pilot study to detect and smooth out any problems.

Economize on the number of variables collected. Only collect the information that is absolutely essential to your study. Use short forms of measurement instruments where possible. Eliminate vague or ambivalent questionnaire items. Use an attractive layout of the instruments. Refrain from using blocks of items that force the respondent to stay on a particular page for a long time. Use computerized adaptive testing where feasible. Do not allow other studies to piggy-back on your data collection efforts.

Do not overdo it. Many Internet questionnaires are annoying because they force the respondent to answer. Do not force your respondent. The result will be an apparently complete dataset with mediocre data. Respect the wish of your respondent to skip items. The end result will be more informative.

Use double coding in the data entry, and chase up any differences between the versions. Devise nonresponse forms in which you try to find out why people they did not respond, or why they dropped out.

Last but not least, consult experts. Many academic centers have departments that specialize in research methodology. Sound expert advice may turn out to be extremely valuable for keeping your missing data rate under control.

Most of this advice can be found in books on research methodology and data quality. Good books are Shadish et al. (2001), De Leeuw et al. (2008), Dillman et al. (2008) and Groves et al. (2009).

1.6.2 Weighting procedures

Weighting is a method to reduce bias when the probability to be selected in the survey differs between respondents. In sample surveys, the responders

are weighted by design weights, which are inversely proportional to their probability of being selected in the survey. If there are missing data, the complete cases are re-weighted according to design weights that are adjusted to counter any selection effects produced by nonresponse. The method is widely used in official statistics. Relevant pointers include Cochran (1977) and Särndal et al. (1992) and Bethlehem (2002).

The method is relatively simple in that only one set of weights is needed for all incomplete variables. On the other hand, it discards data by listwise deletion, and it cannot handle partial response. Expressions for the variance of regression weights or correlations tend to be complex, or do not exist. The weights are estimated from the data, but are generally treated as fixed. The implications for this are unclear (Little and Rubin, 2002, p. 53).

There has been interest recently in improved weighting procedures that are "double robust" (Scharfstein et al., 1999; Bang and Robins, 2005). This estimation method requires specification of three models: Model A is the scientifically interesting model, Model B is the response model for the outcome and model C is the joint model for the predictors and the outcome. The dual robustness property states that: if either Model B or Model C is wrong (but not both), the estimates under Model A are still consistent. This seems like a useful property, but the issue is not free of controversy (Kang and Schafer, 2007).

1.6.3 Likelihood-based approaches

Likelihood-based approaches define a model for the observed data. Since the model is specialized to the observed values, there is no need to impute missing data or to discard incomplete cases. The inferences are based on the likelihood or posterior distribution under the posited model. The parameters are estimated by maximum likelihood, the EM algorithm, the sweep operator, Newton–Raphson, Bayesian simulation and variants thereof. These methods are smart ways to skip over the missing data, and are known as direct likelihood and full information maximum likelihood (FIML).

Likelihood-based methods are, in some sense, the "royal way" to treat missing data problems. The estimated parameters nicely summarize the available information under the assumed models for the complete data and the missing data. The model assumptions can be displayed and evaluated, and in many cases it is possible to estimate the standard error of the estimates.

Multiple imputation is an extension of likelihood-based methods. It adds an extra step in which imputed data values are drawn. An advantage of this is that it is generally easier to calculate the standard errors for a wider range of parameters. Moreover, the imputed values created by multiple imputation can be inspected and analyzed, which helps us to gauge the effect of the model assumptions on the inferences.

The likelihood-based approach receives excellent treatment in the book by Little and Rubin (2002). A less technical account that should appeal to

social scientists can be found in Enders (2010, chapters 3–5). Molenberghs and Kenward (2007) provide a hands-on approach of likelihood-based methods geared toward clinical studies, including extensions to data that are MNAR.

1.7 Structure of the book

This book consists of three main parts: basics, case studies and extensions. Chapter 2 reviews the history of multiple imputation and introduces the notation and theory. Chapter 3 provides an overview of imputation methods for univariate missing data. Chapter 4 distinguishes three approaches to attack the problem of multivariate missing data. Chapter 5 discusses issues that may arise when applying multiple imputation to multivariate missing data. Chapter 6 reviews issues pertaining to the analysis of the imputed datasets.

Chapters 7–9 contain case studies of the techniques described in the previous chapters. Chapter 7 deals with "problems with the columns," while Chapter 8 addresses "problems with the rows." Chapter 9 discusses studies on problems with both rows and columns.

Chapter 10 concludes the main text with a discussion of limitations and pitfalls, reporting guidelines, alternative applications and future extensions. The appendix discusses software options for multiple imputation.

1.8 Exercises

1. *Reporting practice*. What are the reporting practices in your field? Take a random sample of articles that have appeared during the last 10 years in the leading journal in your field. Select only those that present quantitative analyses, and address the following topics:
   (a) Did the authors report that there were missing data?
   (b) If not, can you infer from the text that there must have been missing data?
   (c) Did the authors discuss how they handled the missing data?
   (d) Were the missing data properly addressed?
   (e) Can you detect a trend over time in reporting practice?
   (f) Would the editors of the journal be interested in your findings?

2. *Loss of information*. Suppose that a dataset consists of 100 cases and 10

variables. Each variable contains 10% missing values. What is the largest possible subsample under listwise deletion? What is the smallest? If each variable is MCAR, how many cases will remain?

3. *Stochastic regression imputation.* The correlation of the data in Figure 1.3 is equal to 0.28. Note that this is the lowest of all correlations reported in Section 1.3, which seems to contradict the statement that stochastic regression imputation does not bias the correlation.

   (a) Provide a hypothesis why this correlation is so low.
   (b) Rerun the code with a different `seed` value. What is the correlation now?
   (c) Write a loop to apply apply stochastic regression imputation with the `seed` increasing from 1 to 1000. Calculate the regression weight and the correlation for each solution, and plot the histogram. What are the mean, minimum and maximum values of the correlation?
   (d) Do your results indicate that stochastic regression imputation alters the correlation?

4. *Stochastic regression imputation (continued).* The largest correlation found in the previous exercise exceeds the value found in Section 1.3.4. This seems odd since the correlation of the imputed values under regression imputation is equal to 1, and hence the imputed data have a maximal contribution to the overall correlation.

   (a) Can you explain why this could happen?
   (b) Adapt the code from the previous exercise to test your explanation. Was your explanation satisfactory?
   (c) If not, can you think of another reason, and test that? Hint: Find out what is special about the solutions with the largest correlations.

5. *Nonlinear model.* The model fitted to the `airquality` data in Section 1.4.3 is a simple linear model. Inspection of the residuals reveals that there is a slight curvature in the average of the residuals.

   (a) Start from the completed cases, and use `plot(fit)` to obtain diagnostic plots. Can you explain why the curvature shows up?
   (b) Experiment with solutions, e.g., by transforming `Ozone` or by adding a quadratic term to the model. Can you make the curvature disappear? Does the amount of explained variance increase?
   (c) Does the curvature also show up in the imputed data? If so, does the same solution work? Hint: You can assess the jth fitted model by `fit$analyses[[j]]`, where `fit` was created by `with(imp,...)`.
   (d) Advanced: Do you think your solution would necessitate drawing new imputations?

Chapter 2

Multiple imputation

2.1 Historic overview

2.1.1 Imputation

The English verb "to impute" comes from the Latin *imputo*, which means to reckon, attribute, make account of, charge, ascribe. In the Bible, the word "impute" is a translation of the Hebrew verb *hāshab*, which appears about 120 times in the Old Testament in various meanings (Renn, 2005). The noun "imputation" has a long history in taxation. The concept "imputed income" was used in the 19th century to denote income derived from property, such as land and housing. In the statistical literature, imputation means "filling in the data." Imputation in this sense is first mentioned in 1957 in the work of the U.S. Census Bureau (US Bureau of the Census, 1957).

Allan and Wishart (1930) were the first to develop a statistical method to replace a missing value. They provided two formulae for estimating the value of a single missing observation, and advised filling in the estimate in the data. They would then proceed as usual, but deduct one degree of freedom to correct for the missing data. Yates (1933) generalized this work to more than one missing observation, and thus planted the seeds via a long and fruitful chain of intermediates that led up to the now classic EM algorithm (Dempster et al., 1977). Interestingly, the term "imputation" was not used by Dempster et al. or by any of their predecessors; it only gained widespread use after the monumental work of the Panel on Incomplete Data in 1983. Volume 2 devoted about 150 pages to an overview of the state-of-the-art of imputation technology (Madow et al., 1983). This work is not widely known, but it was the predecessor to the first edition of Little and Rubin (1987), a book that established the term firmly in the mainstream statistical literature.

2.1.2 Multiple imputation

Multiple imputation is now accepted as the best general method to deal with incomplete data in many fields, but this was not always the case. Multiple imputation was developed by Donald B. Rubin in the 1970's. It is useful to know a bit of its remarkable history, as some of the issues in multiple

imputation may resurface in contemporary applications. This section details historical observations that provide the necessary background.

The birth of multiple imputation has been documented by Fritz Scheuren (Scheuren, 2005). Multiple imputation was developed as a solution to a practical problem with missing income data in the March Income Supplement to the Current Population Survey (CPS). In 1977, Scheuren was working on a joint project of the Social Security Administration and the US Census Bureau. The Census Bureau was then using (and still does use) a *hot deck* imputation procedure. Scheuren signaled that the variance could not be properly calculated, and asked Rubin what might be done instead. Rubin came up with the idea of using multiple versions of the complete dataset, something he had already explored in the early 1970s (Rubin, 1994). The original 1977 report introducing the idea was published in 2004 in the history corner of the *American Statistician* (Rubin, 2004). According to Scheuren: "The paper is the beginning point of a truly revolutionary change in our thinking on the topic of missingness" (Scheuren, 2004, p. 291).

Rubin observed that imputing *one* value (single imputation) for the missing value could not be correct in general. He needed a model to relate the unobserved data to the observed data, and noted that even for a given model the imputed values could not be calculated with certainty. His solution was simple and brilliant: create multiple imputations that reflect the uncertainty of the missing data. The 1977 report explains how to choose the models and how to derive the imputations. A low number of imputations, say five, would be enough.

The idea to create multiple versions of the data must have seemed outrageous at that time. Drawing imputations from a distribution, instead of estimating the "best" value, was a drastic departure from everything that had been done before. Rubin's original proposal did not include formulae for calculating combined estimates, but instead stressed the study of variation because of uncertainty in the imputed values. The idea was rooted in the Bayesian framework for inference, quite different from the dominant randomization-based framework in survey statistics. Moreover, there were practical issues involved in the technique, the larger datasets, the extra works to create the model and the repeated analysis, software issues, and so on. These issues have all been addressed by now, but in 1983 Dempster and Rubin wrote: "Practical implementation is still in the developmental state" (Dempster and Rubin, 1983, p. 8).

Rubin (1987a) provided the methodological and statistical footing for the method. Though several improvements have been made since 1987, the book was really ahead of its time and discusses the essentials of modern imputation technology. It provides the formulas needed to combine the repeated complete-data estimates (now called Rubin's rules), and outlines the conditions under which statistical inference under multiple imputation will be valid. Furthermore, pp. 166–170 provide a description of Bayesian sampling algorithms that could be used in practice.

Tests for combinations of parameters were developed by Li et al. (1991a), Li et al. (1991b) and Meng and Rubin (1992). Technical improvements for the degrees of freedom were suggested by Barnard and Rubin (1999) and Reiter (2007). Iterative algorithms for multivariate missing data with general missing data patterns were proposed by Rubin (1987, p. 192) Schafer (1997), Van Buuren et al. (1999), Raghunathan et al. (2001) and King et al. (2001). Additional work on the choice of the number of imputations was done by Royston et al. (2004), Graham et al. (2007) and Bodner (2008).

In the 1990s, multiple imputation came under fire from various sides. The most severe criticism was voiced by Fay (1992). Fay pointed out that the validity of multiple imputation can depend on the form of subsequent analysis. He produced "counterexamples" in which multiple imputation systematically understated the true covariance, and concluded that "multiple imputation is inappropriate as a general purpose methodology." Meng (1994) pointed out that Fay's imputation models omitted important relations that were needed in the analysis model, an undesirable situation that he labeled *uncongenial*. Related issues on the interplay between the imputation model and the complete-data model have been discussed by Rubin (1996) and Schafer (2003).

Several authors have shown that Rubin's estimate of the variance is biased (Wang and Robins, 1998; Robins and Wang, 2000; Nielsen, 2003; Kim et al., 2006). If there is bias, the estimate is usually too large. In response, Rubin (2003) emphasized that variance estimation is only an intermediate goal for making confidence intervals, and that the observed bias does not seem to affect the coverage of these intervals across a wide range of cases of practical interest. He reasoned therefore that these findings do not invalidate multiple imputation in general.

The tide turned around 2005. Reviews that criticize insufficient reporting practice of missing data started to appear in diverse fields (cf. Section 1.1.2). Nowadays multiple imputation is almost universally accepted, and in fact acts as the benchmark against which newer methods are being compared. The major statistical packages have all implemented modules for multiple imputation, so effectively the technology is implemented, almost three decades after Dempster and Rubin's remark.

2.1.3 The expanding literature on multiple imputation

Figure 2.1 contains three time series with counts on the number of publications on multiple imputation during the period 1977–2010. Counts were made in three ways. The rightmost series corresponds to the number of publications per year that featured the search term "multiple imputation" in the title. The search was done in Scopus on July 11, 2011. These are often methodological articles in which new adaptations are being developed. The series in the middle is the number of publication that featured "multiple imputation" in the title, abstract or key words in Scopus on the same search data. This set includes a growing group of papers that contain applications. Scopus does not

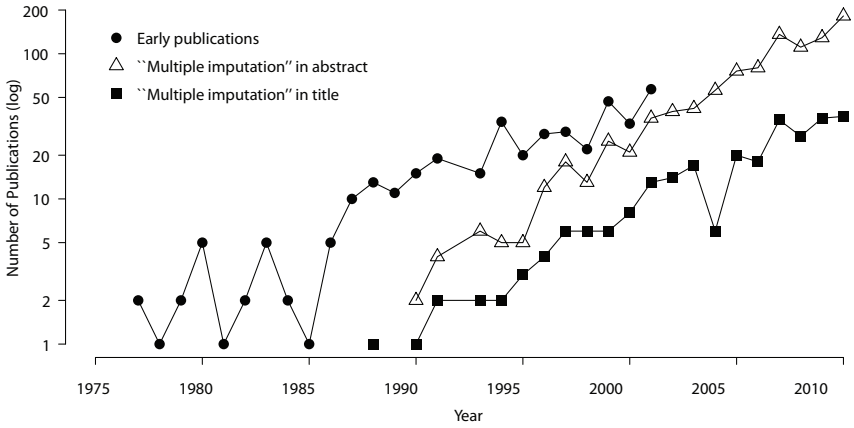


Figure 2.1: Number of publications (log) on multiple imputation during the period 1977–2010 according to three counting methods. Data source: www.scopus.com.

go back further than 1988 on this topic. The leftmost series is the number of publications in a collection of early publications available at www.multiple-imputation.com. This collection covers essentially everything related to multiple imputation from its inception in 1977 up to the year 2001. This group also includes chapters in books, dissertations, conference proceedings, technical reports and so on.

Note that the vertical axis is set in the logarithm. Perhaps the most interesting series is the middle series counting the applications. The pattern is approximately linear, meaning that the number of applications is growing at an exponential rate.

2.2 Concepts in incomplete data

2.2.1 Incomplete data perspective

Many statistical techniques address some kind of incomplete data problem. Suppose that we are interested in knowing the mean income Q in a given population. If we take a sample from the population, then the units not in the sample will have missing values because they will not be measured. It is not possible to calculate the population mean right away since the mean is undefined if one or more values are missing. The incomplete data perspective is a conceptual framework for analyzing data as a missing data problem.

Table 2.1: Examples of reasons for missingness for combinations of intentional/unintentional missing data with item/unit nonresponse

	Intentional	Unintentional
Unit nonresponse	Sampling	Refusal
		Self-selection
Item nonresponse	Matrix sampling	Skip question
	Branching	Coding error

Estimating a mean from a population is a well known problem that can also be solved without a reference to missing data. It is nevertheless sometimes useful to think what we would have done had the data been complete, and what we could do to arrive at complete data. The incomplete data perspective is general, and covers the sampling problem, the counterfactual model of causal inference, statistical modeling of the missing data, and statistical computation techniques. The books by Gelman et al. (2004, ch. 7) and Gelman and Meng (2004) provide in-depth discussions of the generality and richness of the incomplete data perspective.

2.2.2 Causes of missing data

There is a broad distinction between two types of missing data: *intentional* and *unintentional* missing data. Intentional missing data are planned by the data collector. For example, the data of a unit can be missing because the unit was excluded from the sample. Another form of intentional missing data is the use of different versions of the same instrument for different subgroups, an approach known as matrix sampling. See Gonzalez and Eltinge (2007) for an overview. Also, missing data that occur because of the routing in a questionnaire are intentional, as well as data (e.g., survival times) that are censored data at some time because the event (e.g., death) has not yet taken place.

Though often foreseen, unintentional missing data are unplanned and not under the control of the data collector. Examples are: the respondent skipped an item, there was an error in the data transmission causing data to be missing, some of the objects dropped out before the study could be completed resulting in partially complete data, and the respondent was sampled but refused to cooperate.

Another important distinction is *item nonresponse* versus *unit nonresponse*. Item nonresponse refers to the situation in which the respondent skipped one or more items in the survey. Unit nonresponse occurs if the respondent refused to participate, so all outcome data are missing for this respondent. Historically, the methods for item and unit nonresponse have been rather different, with unit nonresponse primarily addressed by weighting methods, and item nonresponse primarily addressed by edit and imputation techniques.

Table 2.1 cross-classifies both distinctions, and provides some typical examples in each of the four cells. The distinction between intentional/unintentional missing data is the more important one conceptually. The item/unit nonresponse distinction says *how much* information is missing, while the distinction between intentional and unintentional missing data says *why* some information is missing. Knowing the reasons why data are incomplete is a first step toward the solution.

2.2.3 Notation

The notation used in this book will be close to that of Rubin (1987a) and Schafer (1997), but there are some exceptions. The symbol m is used to indicate the number of multiple imputations. Compared to Rubin (1987a) the subscript m is dropped from most of the symbols. In Rubin (1987a), Y and R represent the data of the population, whereas in this book Y refers to data of the sample, similar to Schafer (1997). Rubin (1987a) uses X to represent the completely observed covariates in the population. Here we assume that the covariates are possibly part of Y, so there is not always a symbolic distinction between complete covariates and incomplete data. The symbol X is used to indicate the set of predictors in various types of models.

Let Y denote the $n \times p$ matrix containing the data values on p variables for all n units in the sample. We define the *response indicator* R as an $n \times p$ 0–1 matrix. The elements of Y and R are denoted by y_{ij} and r_{ij}, respectively, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. If y_{ij} is observed, then $r_{ij} = 1$, and if y_{ij} is missing, then $r_{ij} = 0$. Note that R is completely observed in the sample. The observed data are collectively denoted by Y_{obs}. The missing data are collectively denoted as Y_{mis}, and contain all elements y_{ij} where $r_{ij} = 0$. When taken together $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ contain the complete data values. However, the values of the part Y_{mis} are unknown to us, and the observed data are thus incomplete. R indicates the values that are masked.

If $Y = Y_{\text{obs}}$ (i.e., if the sample data are completely observed) and if we know the mechanism of how the sample was created, then it is possible to make a valid estimate of the population quantities of interest. For a simple random sample, we could just take the sample mean $\hat{Q}$ as an unbiased estimate of the population mean Q. We will assume throughout this book that we know how to do the correct statistical analysis on the complete data Y. If we cannot do this, then there is little hope that we can solve the more complex problem of analyzing Y_{obs}. This book addresses the problem of what to do if Y is observed incompletely. Incompleteness can incorporate intentional missing data, but also unintentional forms like refusals, self-selection, skipped questions, missed visits and so on.

Note that every unit in the sample has a row in Y. If no data have been obtained for a unit i (presumably because of unit nonresponse), the ith record will contain only the sample number and perhaps administrative data from the sampling frame. The remainder of the record will be missing.

2.2.4 MCAR, MAR and MNAR again

Section 1.2 introduced MCAR, MAR and MNAR. This section provides more precise definitions.

The matrix R stores the locations of the missing data in Y. The distribution of R may depend on $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, either by design or by happenstance, and this relation is described by the *missing data model*. Let ψ contain the parameters of the missing data model, then the general expression of the missing data model is $\Pr(R|Y_{\text{obs}}, Y_{\text{mis}}, \psi)$.

The data are said to be MCAR if

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|\psi) \tag{2.1}$$

so the probability of being missing depends only on some parameters ψ, the overall probability of being missing. The data are said to be MAR if

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|Y_{\text{obs}}, \psi) \tag{2.2}$$

so the missingness probability may depend on observed information, including any design factors. Finally, the data are MNAR if

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) \tag{2.3}$$

does not simplify, so here the probability to be missing also depends on unobserved information, including Y_{mis} itself.

As explained in Chapter 1, simple techniques usually only work under MCAR, but this assumption is very restrictive and often unrealistic. Multiple imputation can handle both MAR and MNAR.

Several tests have been proposed to test MCAR versus MAR. These tests are not widely used, and their practical value is unclear. See Enders (2010, pp. 17–21) for an evaluation of two procedures. It is not possible to test MAR versus MNAR since the information that is needed for such a test is missing.

Numerical illustration. We simulate three archetypes of MCAR, MAR and MNAR. The data $Y = (Y_1, Y_2)$ are drawn from a standard bivariate normal distribution with a correlation between Y_1 and Y_2 equal to 0.5. Missing data are created in Y_2 using the missing data model

$$\Pr(R_2 = 0) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}}\psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}}\psi_2 \tag{2.4}$$

with different parameters settings for $\psi = (\psi_0, \psi_1, \psi_2)$. For MCAR we set $\psi_{\text{MCAR}} = (0.5, 0, 0)$, for MAR we set $\psi_{\text{MAR}} = (0, 1, 0)$ and for MNAR we set $\psi_{\text{MNAR}} = (0, 0, 1)$. Thus, we obtain the following models:

$$\text{MCAR} \quad : \quad \Pr(R_2 = 0) = 0.5 \tag{2.5}$$

$$\text{MAR} \quad : \quad \text{logit}(\Pr(R_2 = 0)) = Y_1 \tag{2.6}$$

$$\text{MNAR} \quad : \quad \text{logit}(\Pr(R_2 = 0)) = Y_2 \tag{2.7}$$

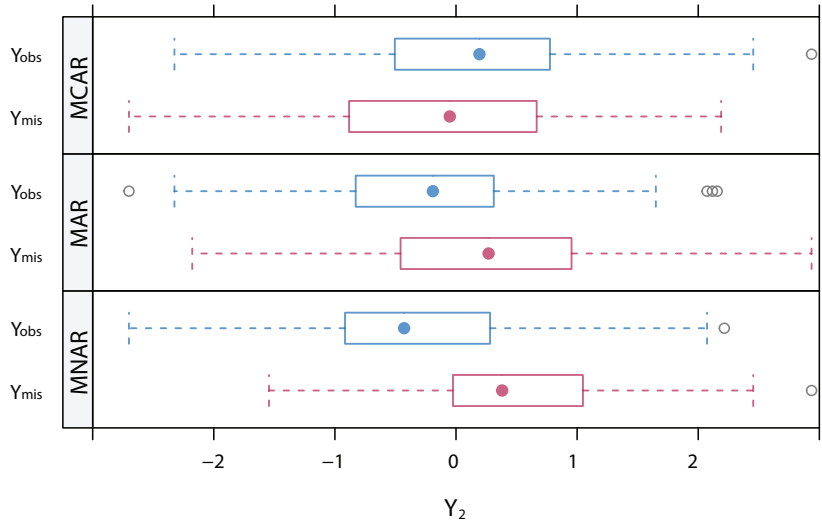


Figure 2.2: Distribution of Y_{obs} and Y_{mis} under three missing data models.

where $\text{logit}(p) = \log(p/(1-p))$ for any $0 < p < 1$ is the logit function. In practice, it is more convenient to work with the inverse logit (or logistic) function $\text{inverse logit}^{-1}(x) = \exp(x)/(1+\exp(x))$, which transforms a continuous x to the interval $\langle 0, 1\rangle$. In `R`, it is straightforward to draw random values under these models as

```
> library(MASS)
> logistic <- function(x) exp(x)/(1 + exp(x))
> set.seed(80122)
> n <- 300
> y <- mvrnorm(n = n, mu = c(0, 0), Sigma = matrix(c(1,
      0.5, 0.5, 1), nrow = 2))
> y1 <- y[, 1]
> y2 <- y[, 2]
> r2.mcar <- 1 - rbinom(n, 1, 0.5)
> r2.mar <- 1 - rbinom(n, 1, logistic(y1))
> r2.mnar <- 1 - rbinom(n, 1, logistic(y2))
```

Figure 2.2 displays the distribution of Y_{obs} and Y_{mis} under the three missing data models. As expected, these are similar under MCAR, but become progressively more distinct as we move to the MNAR model.

2.2.5 Ignorable and nonignorable ♠

The example in the preceding section specified parameters ψ for three missing data models. The ψ-parameters have no intrinsic scientific value and are generally unknown. It would simplify the analysis if we could just ignore these parameters. The practical importance of the distinction between MCAR, MAR and MNAR is that it clarifies the conditions under which we can accurately estimate the scientifically interesting parameters without the need to know ψ.

The actually observed data consist of Y_{obs} and R. The joint density function $f(Y_{\text{obs}}, R|\theta, \psi)$ of Y_{obs} and R together depends on parameters θ for the full data Y that are of scientific interest, and parameters ψ for the response indicator R that are seldom of interest. The joint density is proportional to the likelihood of θ and ψ, i.e.,

$$l(\theta, \psi|Y_{\text{obs}}, R) \propto f(Y_{\text{obs}}, R|\theta, \psi) \tag{2.8}$$

The question is: When can we determine θ without knowing ψ, or equivalently, the mechanism that created the missing data? The answer is given in Little and Rubin (2002, p. 119):

> The missing data mechanism is ignorable for likelihood inference if:
>
> 1. MAR: the missing data are missing at random; and
> 2. Distinctness: the parameters θ and ψ are distinct, in the sense that the joint parameter space of (ψ, θ) is the product of the parameter space of θ and the parameter space of ψ.

For valid Bayesian inference, the latter condition is slightly stricter: θ and ψ should be a priori independent: $p(\theta, \psi) = p(\theta)p(\psi)$ (Little and Rubin, 2002, p. 120). The MAR requirement is generally considered to be the more important condition. Schafer (1997, p. 11) says that in many situations the condition on the parameters is "intuitively reasonable, as knowing θ will provide little information about ψ and vice-versa." We should perhaps be careful in situations where the scientific interest focuses on the missing data process itself. For all practical purposes, the missing data model is said to be "ignorable" if MAR holds.

Note that the label "ignorable" does not mean that we can be entirely careless about the missing data. For inferences to be valid, we need to condition on those factors that influence the missing data rate. For example, in the MAR example of Section 2.2.4 the missingness in Y_2 depends on Y_1. A valid estimate of the mean of Y_2 cannot be made without Y_1, so we should include Y_1 somehow into the calculations for the mean of Y_2.

2.2.6 Implications of ignorability

The concept of ignorability plays an important role in the construction of imputation models. In imputation, we want to draw synthetic observations from the posterior distribution of the missing data, given the observed data and given the process that generated the missing data. The distribution is denoted as $P(Y_{\text{mis}}|Y_{\text{obs}}, R)$. If the nonresponse is ignorable, then this distribution does not depend on R (Rubin, 1987a, Result 2.3), i.e.,

$$P(Y_{\text{mis}}|Y_{\text{obs}}, R) = P(Y_{\text{mis}}|Y_{\text{obs}}) \tag{2.9}$$

The implication is that

$$P(Y|Y_{\text{obs}}, R = 1) = P(Y|Y_{\text{obs}}, R = 0) \tag{2.10}$$

so the distribution of the data Y is the same in the response and nonresponse groups. Thus, if the missing data model is ignorable we can model the posterior distribution $P(Y|Y_{\text{obs}}, R = 1)$ from the observed data, and use this model to create imputations for the missing data. Vice versa, techniques that (implicitly) assume equivalent distributions assume ignorability and thus MAR. On the other hand, if the nonresponse is nonignorable, we find

$$P(Y|Y_{\text{obs}}, R = 1) \neq P(Y|Y_{\text{obs}}, R = 0) \tag{2.11}$$

so then we should incorporate R into the model to create imputations.

The assumption of ignorability is often sensible in practice, and generally provides a natural starting point. If, on the other hand, the assumption is not reasonable (e.g., when data are censored), we may specify $P(Y|Y_{\text{obs}}, R = 0)$ different from $P(Y|Y_{\text{obs}}, R = 1)$. The specification of $P(Y|Y_{\text{obs}}, R = 0)$ needs assumptions external to the data since, by definition, the information needed to estimate any regression weights for R is missing.

Example. Suppose that a growth study measures body weight in kg (Y_2) and gender (Y_1: 1 = boy, 0 = girl) of 15-year-old children, and that some of the body weights are missing. We can model the weight distribution for boys and girls separately for those with observed weights, i.e., $P(Y_2|Y_1 = 1, R_2 = 1)$ and $P(Y_2|Y_1 = 0, R_2 = 1)$. If we assume that the response mechanism is ignorable, then imputations for a boy's weight can be drawn from $P(Y_2|Y_1 = 1, R_2 = 1)$ since it will equal $P(Y_2|Y_1 = 1, R_2 = 0)$. The same can be done for the girls. This procedure leads to correct inferences on the combined sample of boys and girls, even if boys have substantially more missing values, or if the body weights of the boys and girls are very different.

The procedure outlined above is not appropriate if, within the boys or the girls, the occurrence of the missing data is related to body weight. For example, some of the heavier children may not want to be weighed, resulting in more missing values for the obese. It will be clear that assuming $P(Y_2|Y_1, R_2 = 0) = P(Y_2|Y_1, R_2 = 1)$ will underestimate the prevalence of overweight and obesity. In this case, it may be more realistic to specify $P(Y_2|Y_1, R_2 = 0)$ such that

imputation accounts for the excess body weights in the children that were not weighed. There are many ways to do this. In all these cases the response mechanism will be nonignorable.

The assumption of ignorability is essentially the belief on the part of the user that the available data are sufficient to correct for the effects of the missing data. The assumption cannot be tested on the data itself, but it can be checked against suitable external validation data.

There are two main strategies that we may pursue if the response mechanism is not ignorable. The first is to expand the data, and assume ignorability on the expanded data (Collins et al., 2001). See also Section 5.2 for more details. In the above example, overweight children may simply not want anybody to know their weight, but perhaps have no objection if their waist circumference Y_3 is measured. As Y_3 predicts Y_2, R_2 or both, the ignorability assumption $P(Y_2|Y_1, Y_3, R_2 = 0) = P(Y_2|Y_1, Y_3, R_2 = 1)$ is less stringent, and hence more realistic.

The second strategy is to formulate the model for $P(Y_2|Y_1, R_2 = 0)$ different from $P(Y_2|Y_1, R_2 = 1)$, describing which body weights would have been observed if they had been measured. Such a model could simply add some extra kilos, known as δ-adjustment, to the imputed values, but of course we need to be able to justify our choice in light of what we know about the data. See Section 3.9.1 for a more detailed discussion of the idea. In general, the formulation of nonignorable models should be driven by knowledge about the process that created the missing data. Any such methods need to be explained and justified as part of the statistical analysis.

2.3 Why and when multiple imputation works

2.3.1 Goal of multiple imputation

A *scientific estimand* Q is a quantity of scientific interest that we can calculate if we would observe the entire population. For example, we could be interested in the mean income of the population. In general, Q can be expressed as a known function of the population data. If we are interested in more than one quantity, Q will be a vector. Note that Q is a property of the population, so it does not depend on any design characteristics. Examples of scientific estimands include the population mean, the population (co)variance or correlation, and population factor loadings and regression coefficients, as well as these quantities calculated within known strata of the population. Examples of quantities that are not scientific estimands are sample means, standard errors and test statistics.

We can only calculate Q if the population data are fully known, but this is almost never the case. The goal of multiple imputation is to find an *estimate* $\hat{Q}$

that is *unbiased* and *confidence valid* (Rubin, 1996). We explain these concepts below.

Unbiasedness means that the average $\hat{Q}$ over all possible samples Y from the population is equal to Q. The formula is

$$E(\hat{Q}|Y) = Q \tag{2.12}$$

The explanation of confidence validity requires some additional symbols. Let U be the estimated variance-covariance matrix of $\hat{Q}$. This estimate is *confidence valid* if the average of U over all possible samples is equal or larger than the variance of $\hat{Q}$. The formula is

$$E(U|Y) \geq V(\hat{Q}|Y) \tag{2.13}$$

where the function $V(\hat{Q}|Y)$ denotes the variance caused by the sampling process. A statistical test with a stated nominal rejection rate of 5% should reject the null hypothesis in at most 5% of the cases when in fact the null hypothesis is true. A procedure is said to be confidence valid if this holds.

In summary, the goal of multiple imputation is to obtain estimates of the scientific estimand in the population. This estimate should on average be equal to the value of the population parameter. Moreover, the associated confidence intervals and hypothesis tests should achieve at least the stated nominal value.

2.3.2 Three sources of variation ♠

The actual value of Q is unknown if some of the population data are unknown. Suppose we make an estimate $\hat{Q}$ of Q. The amount of uncertainty in $\hat{Q}$ about the true population value Q depends on what we know about Y_{mis}. If we would be able to re-create Y_{mis} perfectly, then we can calculate Q with certainty. However, such perfect re-creation is almost never unachievable. In other cases, we need to summarize the distribution of Q under varying Y_{mis}.

The possible values of Q given our knowledge of the data Y_{obs} are captured by the posterior distribution $P(Q|Y_{\text{obs}})$. In itself, $P(Q|Y_{\text{obs}})$ is often intractable, but it can be decomposed into two parts that are easier to solve as follows:

$$P(Q|Y_{\text{obs}}) = \int P(Q|Y_{\text{obs}}, Y_{\text{mis}})P(Y_{\text{mis}}|Y_{\text{obs}})dY_{\text{mis}} \tag{2.14}$$

Here, $P(Q|Y_{\text{obs}})$ is the posterior distribution of Q given the observed data Y_{obs}. This is the distribution that we would like to know. $P(Q|Y_{\text{obs}}, Y_{\text{mis}})$ is the posterior distribution of Q in the hypothetically complete data, and $P(Y_{\text{mis}}|Y_{\text{obs}})$ is the posterior distribution of the missing data given the observed data.

The interpretation of Equation 2.14 is most conveniently done from right to left. Suppose that we use $P(Y_{\text{mis}}|Y_{\text{obs}})$ to draw imputations for Y_{mis}, denoted as $\dot{Y}_{\text{mis}}$. We can then use $P(Q|Y_{\text{obs}}, \dot{Y}_{\text{mis}})$ to calculate the quantity of interest Q from the hypothetically complete data $(Y_{\text{obs}},\dot{Y}_{\text{mis}})$. We repeat these two steps

with new draws $\dot{Y}_{\text{mis}}$, and so on. Equation 2.14 says that the actual posterior distribution of Q is equal to the average over the repeated draws of Q. This result is important since it expresses $P(Q|Y_{\text{obs}})$, which is generally difficult, as a combination of two simpler posteriors from which draws can be made.

It can be shown that the posterior mean of $P(Q|Y_{\text{obs}})$ is equal to

$$E(Q|Y_{\text{obs}}) = E(E[Q|Y_{\text{obs}}, Y_{\text{mis}}]|Y_{\text{obs}}) \tag{2.15}$$

the average of the posterior means of Q over the repeatedly imputed data. This equation suggests the following procedure for combining the results of repeated imputations. Suppose that $\hat{Q}_l$ is the estimate of the ℓth repeated imputation, then the combined estimate is equal to

$$\bar{Q} = \frac{1}{m}\sum_{\ell=1}^{m}\hat{Q}_\ell \tag{2.16}$$

where $\hat{Q}_\ell$ contains k parameters and is represented as a $k \times 1$ column vector.

The posterior variance of $P(Q|Y_{\text{obs}})$ is the sum of two variance components:

$$V(Q|Y_{\text{obs}}) = E[V(Q|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}] + V[E(Q|Y_{\text{obs}}, Y_{\text{mis}})|Y_{\text{obs}}] \tag{2.17}$$

This equation is well known in statistics, but can be difficult to grasp at first. The first component is the average of the repeated complete data posterior variances of Q. This is called the within-variance. The second component is the variance between the complete data posterior means of Q. This is called the between variance. Let $\bar{U}_\infty$ and B_∞ denote the estimated within and between components for an infinitely large number of imputations $m = \infty$. Then $T_\infty = \bar{U}_\infty + B_\infty$ is the posterior variance of Q.

Equation 2.17 suggests the following procedure to estimate T_∞ for finite m. We calculate the average of the complete-data variances as

$$\bar{U} = \frac{1}{m}\sum_{\ell=1}^{m}\bar{U}_\ell \tag{2.18}$$

where the term $\bar{U}_\ell$ is the variance-covariance matrix of $\hat{Q}_\ell$ obtained for the ℓth imputation. The standard unbiased estimate of the variance between the m complete data estimates is given by

$$B = \frac{1}{m-1}\sum_{\ell=1}^{m}(\hat{Q}_\ell - \bar{Q})(\hat{Q}_\ell - \bar{Q})' \tag{2.19}$$

where $\bar{Q}$ is calculated by Equation 2.16.

It is tempting to conclude that the total variance T is equal to the sum of $\bar{U}$ and B, but that would be incorrect. We need to incorporate the fact that $\bar{Q}$ itself is estimated using finite m, and thus only approximates $\bar{Q}_\infty$. Rubin

(1987a, eq. 3.3.5) shows that the contribution to the variance of this factor is systematic and equal to B_∞/m. Since B approximates B_∞, we may write

$$\begin{aligned} T &= \bar{U} + B + B/m \\ &= \bar{U} + \left(1 + \frac{1}{m}\right) B \end{aligned} \tag{2.20}$$

for the total variance of $\bar{Q}$, and hence of $(Q-\bar{Q})$ if $\bar{Q}$ is unbiased. The procedure to combine the repeated-imputation results by Equations 2.16 and 2.20 is referred to as Rubin's rules.

In summary, the total variance T stems from three sources:

1. $\bar{U}$, the variance caused by the fact that we are taking a sample rather than observing the entire population. This is the conventional statistical measure of variability;
2. B, the extra variance caused by the fact that there are missing values in the sample;
3. B/m, the extra simulation variance caused by the fact that $\bar{Q}$ itself is estimated for finite m.

The addition of the latter term is critical to make multiple imputation work at low values of m. Not including it would result in p-values that are too low, or confidence intervals that are too short.

Traditional choices for m are $m = 3$, $m = 5$ and $m = 10$. The current advice is to set m higher, e.g., $m = 50$ (cf. Section 2.7). The larger m gets, the smaller the effect of simulation error on the total variance.

2.3.3 Proper imputation

In order to yield valid statistical inferences, the imputed values should possess certain characteristics. Procedures that yield such imputations are called *proper* (Rubin, 1987a, pp. 118–128). Section 2.3.1 described two conditions needed for a valid estimate of Q. These requirements apply simultaneously to both the sampling and the nonresponse model. An analogous set of requirements exists if we zoom in on procedures that deal exclusively with the response model. The important theoretical result is: If the imputation method is proper and if the complete data analysis is valid in the sense of Section 2.3.1, the whole procedure is valid (Rubin, 1987a, p. 119).

Recall from Section 2.3.1 that the goal of multiple imputation is to find an estimate $\hat{Q}$ of Q with correct statistical properties. At the level of the sample, there is uncertainty about Q. This uncertainty is captured by U, the estimated variance-covariance of $\hat{Q}$ in the sample. If we have no missing data in the sample, the pair $(\hat{Q}, U)$ contains everything we know about Q.

If we have incomplete data, we can distinguish three analytic levels: the population, the sample and the incomplete sample. The problem of estimating

Table 2.2: Role of symbols at three analytic levels and the relations between them. The relation $\Longrightarrow$ means "is an estimate of." The relation $\doteq$ means "is asymptotically equal to."

Incomplete Sample Y_{obs}		Complete Sample $Y = (Y_{\text{obs}}, Y_{\text{mis}})$		Population
$\bar{Q}$	$\Longrightarrow$	$\hat{Q}$	$\Longrightarrow$	Q
$\bar{U}$	$\Longrightarrow$	$U \doteq V(\hat{Q})$		
$B \doteq V(\bar{Q})$				

Q in the population by $\hat{Q}$ from the sample is a traditional statistical problem. The key idea of the solution is to accompany $\hat{Q}$ by an estimate of its variability under repeated sampling U according to the sampling model.

Now suppose that we want to go from the incomplete sample to the complete sample. At the sample level we can distinguish two estimands, instead of one: $\hat{Q}$ and U. Thus, the role of the single estimand Q at the population level is taken over by the estimand pair $(\hat{Q}, U)$ at the sample level. Table 2.2 provides an overview of the three different analytic levels involved, the quantities defined at each level and their relations. Note that $\hat{Q}$ is both an estimate (of Q) as well as an estimand (of $\bar{Q}$). Also, U has two roles.

Imputation is the act of converting an incomplete sample into a complete sample. Imputation of data should, at the very least, lead to adequate estimates of both $\hat{Q}$ and U. Three conditions define whether an imputation procedure is considered proper. We use the slightly simplified version given by Brand (1999, p. 89) combined with Rubin (1987a). An imputation procedure is said to be *confidence proper* for complete data statistics $(\hat{Q}, U)$ if at large m all of the following conditions hold approximately:

$$E(\bar{Q}|Y) = \hat{Q} \tag{2.21}$$

$$E(\bar{U}|Y) = U \tag{2.22}$$

$$\left(1 + \frac{1}{m}\right) E(B|Y) \geq V(\bar{Q}) \tag{2.23}$$

The hypothetically complete sample data Y is now held fixed, and the response indicator R varies according to a specified model.

The first requirement is that $\bar{Q}$ is an unbiased estimate of $\hat{Q}$. This means that, when averaged over the response indicators R sampled under the assumed response model, the multiple imputation estimate $\bar{Q}$ is equal to $\hat{Q}$, the estimate calculated from the hypothetically complete data.

The second requirement is that $\bar{U}$ is an unbiased estimate of U. This means that, when averaged over the response indicator R sampled under the assumed response model, the estimate $\bar{U}$ of the sampling variance of $\hat{Q}$ is equal to U, the sampling variance estimate calculated from the hypothetically complete data.

The third requirement is that B is a confidence valid estimate of the variance due to missing data. Equation 2.23 implies that the extra inferential uncertainty about $\hat{Q}$ due to missing data is correctly reflected. On average, the estimate B of the variance due to missing data should be equal to $V(\bar{Q})$, the variance observed in the multiple imputation estimator $\bar{Q}$ over different realizations of the response mechanism. This requirement is analogous to Equation 2.13 for confidence valid estimates of U.

If we replace $\geq$ in Equation 2.23 by $>$, then the procedure is said to be *proper*, a stricter version. In practice, being confidence proper is enough to obtain valid inferences.

Note a procedure may be proper for the estimand pair $(\hat{Q}, U)$, while being improper for another pair $(\hat{Q}', U')$. Also, a procedure may be proper with respect to one response mechanism $P(R)$, but improper for an alternative mechanism $P(R')$.

It is not always easy to check whether a certain procedure is proper. Section 2.5 describes simulation-based tools for checking the adequacy of imputations for valid statistical inference. Chapter 3 provides examples of proper and improper procedures.

2.3.4 Scope of the imputation model

Imputation models vary in their scope. Models with a narrow scope are proper with respect to specific estimand $(\hat{Q}, U)$ and particular response mechanism, e.g., a particular proportion of nonresponse. Models with a broad scope are proper with respect to a wide range of estimates $\hat{Q}$, e.g., subgroup means, correlations, ratios and so on, and under a large variety of response mechanisms.

The scope is related to the setting in which the data are collected. The following list distinguishes three typical situations:

- *Broad.* Create one set of imputations to be used for all projects and analyses. A broad scope is appropriate for publicly released data, cohort data and registers, where different people use the data for different purposes.

- *Intermediate.* Create one set of imputations per project and use this set for all analyses. An intermediate scope is appropriate for analyses that estimate relatively similar quantities. The imputer and analyst can be different persons.

- *Narrow.* A separate imputed dataset is created for each analysis. The imputer and analyst are typically the same person. A narrow scope is appropriate if the imputed data are used only to estimate the same quantity. Different analyses require different imputations.

In general, imputations created under a broad scope can be applied more widely, and are hence preferable. On the other hand, if we are reasonably

certain of the correct model for the data, and if the imputed data are only used to fit that model, then using a narrow scope can be more efficient. As the correct model is typically unknown, the techniques discussed in this book will generally attempt to create imputations with a broader scope. Whatever is chosen, it is the responsibility of the imputer to indicate the scope of the generated imputations.

2.3.5 Variance ratios ♠

For scalar Q, the ratio

$$\lambda = \frac{B + B/m}{T} \tag{2.24}$$

can be interpreted as the proportion of the variation attributable to the missing data. It is equal to zero if the missing data do not add extra variation to the sampling variance, an exceptional situation that can occur only if we can perfectly re-create the missing data. The maximum value is equal to 1, which occurs only if all variation is caused by the missing data. This is equally unlikely to occur in practice since it means that there is no information at all. If λ is high, say $\lambda > 0.5$, the influence of the imputation model on the final result is larger than that of the complete data model.

The ratio

$$r = \frac{B + B/m}{\bar{U}} \tag{2.25}$$

is called the *relative increase in variance due to nonresponse* (Rubin, 1987a, eq. 3.1.7). The quantity is related to λ by $r = \lambda/(1 - \lambda)$.

Another related measure is the *fraction of information about Q missing due to nonresponse* (Rubin, 1987a, eq. 3.1.10). This measure is defined by

$$\gamma = \frac{r + 2/(\nu + 3)}{1 + r} \tag{2.26}$$

This measure needs an estimate of the degrees of freedom ν, and will be discussed in Section 2.3.6. The interpretations of γ and λ are similar, but γ is adjusted for the finite number of imputations. Both statistics are related by

$$\gamma = \frac{\nu + 1}{\nu + 3}\lambda + \frac{2}{\nu + 3} \tag{2.27}$$

The literature often confuses γ and λ, and erroneously labels λ as the fraction of missing information. The values of λ and γ are almost identical for large ν, but they could notably differ for low ν.

If Q is a vector, it is sometimes useful to calculate a compromise λ over all elements in $\bar{Q}$ as

$$\bar{\lambda} = \left(1 + \frac{1}{m}\right) \text{tr}(BT^{-1})/k \tag{2.28}$$

where k is the dimension of $\bar{Q}$, and where B and T are now $k \times k$ matrices. The compromise expression for r is equal to

$$\bar{r} = \left(1 + \frac{1}{m}\right) \text{tr}(B\bar{U}^{-1})/k \tag{2.29}$$

the average relative increase in variance.

The quantities λ, r and γ as well as their multivariate analogues $\bar{\lambda}$ and $\bar{r}$ are indicators of the severity of the missing data problem. Fractions of missing information up to 0.2 can be interpreted as "modest," 0.3 as "moderately large" and 0.5 as "high" (Li et al., 1991b). High values indicate a difficult problem in which the final statistical inferences are highly dependent on the way in which the missing data were handled. Note that estimates of λ, r and γ may be quite variable for low m (cf. Section 2.7).

2.3.6 Degrees of freedom ♠

The calculation of the degrees of freedom for statistical tests needs some attention since part of the data is missing. The "old" formula (Rubin, 1987a, eq. 3.1.6) for the degrees of freedom can be written concisely as

$$\begin{aligned} \nu_{\text{old}} &= (m-1)\left(1 + \frac{1}{r^2}\right) \\ &= \frac{m-1}{\lambda^2} \end{aligned} \tag{2.30}$$

with r and λ defined as in Section 2.3.5. The lowest possible value is $\nu_{\text{old}} = m-1$, which occurs if essentially all variation is attributable to the nonresponse. The highest value $\nu_{\text{old}} = \infty$ indicates that all variation is sampling variation, either because there were no missing data, or because we could re-create them perfectly.

Barnard and Rubin (1999) noted that Equation 2.30 can produce values that are larger than the degrees of freedom in the complete data, a situation they considered "clearly inappropriate." They developed an adapted version for small samples that is free of the problem. Let ν_{com} be the degrees of freedom of $\bar{Q}$ in the hypothetically complete data. In models that fit k parameters on data with a sample size of n we may set $\nu_{\text{com}} = n-k$. The estimated observed data degrees of freedom that accounts for the missing information is

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \nu_{\text{com}}(1 - \lambda) \tag{2.31}$$

The adjusted degrees of freedom to be used for testing in multiple imputation can be written concisely as

$$\nu = \frac{\nu_{\text{old}} \nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}} \tag{2.32}$$

The quantity ν is always less than or equal to ν_{com}. If $\nu_{\text{com}} = \infty$, then Equation 2.32 reduces to 2.30. If $\lambda = 0$ then $\nu = \nu_{\text{com}}$, and if $\lambda = 1$ we find $\nu = 0$. Distributions with zero degrees of freedom are nonsensical, so for $\nu < 1$ we should refrain from any testing due to lack of information.

2.3.7 Numerical example

Many quantities introduced in the previous sections can be obtained by the `pool()` function in `mice`. The following code imputes the `nhanes` dataset, fits a simple linear model and pools the results:

```
> library("mice")
> options(digits = 3)
> imp <- mice(nhanes, print = FALSE, m = 10, seed = 24415)
> fit <- with(imp, lm(bmi ~ age))
> est <- pool(fit)
```

The statistics are stored as components of the `est` object. We obtain the names of the components by

```
> names(est)
 [1] "call"   "call1"  "call2"  "nmis"   "m"      "qhat"
 [7] "u"      "qbar"   "ubar"   "b"      "t"      "r"
[13] "dfcom"  "df"     "fmi"    "lambda"
```

It should be obvious from the name to which quantity each refers, but if not, type `?pool` for help. Individual components are obtained by, for example, `est$df`, or by using `attach()`. As an illustration, we recalculate the fraction of missing information γ according to Equations 2.26 and 2.27 as

```
> attach(est)
> (r + 2/(df + 3))/(r + 1)
(Intercept)         age
      0.509       0.419
> (df + 1)/(df + 3) * lambda + 2/(df + 3)
(Intercept)         age
      0.509       0.419
> detach(est)
```

2.4 Statistical intervals and tests

2.4.1 Scalar or multi-parameter inference?

The ultimate objective of multiple imputation is to provide valid statistical estimates from incomplete data. For scalar Q, it is straightforward to calculate confidence intervals and p-values from multiply imputed data, the primary difficulty being the derivation of the appropriate degrees of freedom for the t- and F-distributions. Section 2.4.2 provides the relevant statistical procedures.

If Q is a vector, we have two options for analysis. The first option is to calculate confidence intervals and p-values for the individual elements in Q, and do all statistical tests per element. Such repeated-scalar inference is appropriate if we interpret each element as a separate, though perhaps related, model parameter. In this case, the test uses the fraction of missing information particular to each parameter.

The alternative option is to perform one statistical test that involves the elements of Q at once. This is appropriate in the context of multi-parameter or simultaneous inference, where we evaluate combinations of model parameters. Practical applications of such tests include the comparison of nested models and the testing of model terms that involved multiple parameters like regression estimates for dummy codings created from the same variable.

All methods assume that, under repeated sampling and with complete data, the parameter estimates $\hat{Q}$ are normally distributed around the population value Q as

$$\hat{Q} \sim N(Q, U) \tag{2.33}$$

where U is the variance-covariance matrix of $(Q - \hat{Q})$ (Rubin, 1987a, p. 75). For scalar Q, the quantity U reduces to σ_m^2, the variance of the estimate $\hat{Q}$ over repeated samples. Observe that U is not the variance of the measurements.

Several approaches for multiparameter inference are available: Wald test, likelihood ratio test and χ^2-test. These methods are more complex than single-parameter inference, and their treatment is therefore deferred to Section 6.2. The next section shows how confidence intervals and p-values for scalar parameters can be calculated from multiply imputed data.

2.4.2 Scalar inference

Single parameter inference applies if $k = 1$, or if $k > 1$ and the test is repeated for each of the k components. Since the total variance of T is not known a priori, $\bar{Q}$ follows a t-distribution rather than the normal. Univariate tests are based on the approximation

$$\frac{Q - \bar{Q}}{\sqrt{T}} \sim t_\nu \tag{2.34}$$

where t_ν is the Student's t-distribution with ν degrees of freedom, with ν defined by Equation 2.32.

The $100(1-\alpha)\%$ confidence interval of a $\bar{Q}$ is calculated as

$$\bar{Q} \pm t_{\nu,1-\alpha/2}\sqrt{T} \tag{2.35}$$

where $t_{\nu,1-\alpha/2}$ is the quantile corresponding to probability $1-\alpha/2$ of t_ν. For example, use $t_{10,0.975} = 2.23$ for the 95% confidence interval with $\nu = 10$.

Suppose we test the null hypothesis $Q = Q_0$ for some specified value Q_0. We can find the p-value of the test as the probability

$$P_s = \Pr\left[F_{1,\nu} > \frac{(Q_0 - \bar{Q})^2}{T}\right] \tag{2.36}$$

where $F_{1,\nu}$ is an F-distribution with 1 and ν degrees of freedom.

Statistical intervals and tests based on scalar Q are used most widely, and are typically available from multiple imputation software. In `mice` these tests are standard output of the `pool()` and `summary.mipo()` functions. We have already seen examples of this in Section 1.4.3.

2.5 Evaluation criteria

2.5.1 Imputation is not prediction

The goal of multiple imputation is to obtain statistically valid inferences from incomplete data. Though ambitious, this goal is achievable. Many investigators new to the field of missing data are however trying to achieve an even more ambitious goal: to re-create the lost data. This more ambitious goal is often not stated as such, but it can be diagnosed easily from the simulations presented. Simulations typically start from a complete dataset, generate missing data in some way, and then fill in the missing data by the old and new procedures. In the world of simulation we have access to both the true and imputed values, so the obvious way to compare the old and new methods would be to study how well they can recreate the true data. The method that best recovers the true data "wins."

There is however a big problem with this procedure. The method that best recovers the true data may be nonsensical or may contain severe flaws. Let us look at an example. Suppose we start from a complete dataset of adult male heights, randomly delete 50% of the data, apply our favorite imputation procedure and measure how well it recovers the true data. For example, let the criterion for accuracy be the sum of squared differences

$$\sum_{i=1}^{n_{\text{mis}}} (y_i^{\text{mis}} - \dot{y}_i)^2 \tag{2.37}$$

where y_i^{mis} represents the true (but usually unknown) height of person i, and where $\dot{y}_i$ is imputed height of person i. It is not difficult to show that the method that yields the best recovery of the true values is mean imputation. However, it makes little sense to impute the same height for everyone, since the true missing heights y_i^{mis} vary as much as those in the observed data. As outlined in Section 1.3.3, mean imputation has severe shortcomings: It distorts the height distribution and affect the variance in the imputed data. Thus, an accuracy criterion like Equation 2.37 may look objective, but in fact it promotes an intrinsically flawed imputation method as best. The conclusion must be that minimizing discrepancy between the true data and the imputed data is not a valid way to create imputed datasets.

The problem is this: Optimizing an imputed value with respect to a criterion searches for its most probable value. However, the most likely value may perhaps be observed in 20% of the cases, which means that in 80% of the cases it will not be observed. Using the most probable value as an imputed value in 100% of the cases ignores the uncertainty associated with the missing data. As a consequence, various distortions may result when the completed data are analyzed. We may find an artificial increase in the strength of the relationship between the variables. Also, confidence intervals may be too short. These drawbacks result from treating the problem as a prediction (or optimization) problem that searches for the best value.

The evaluation of accuracy is nevertheless deeply rooted within many fields. Measures of *accuracy* include any statistic that is based on the difference between true and the imputed data. Examples abound in pattern recognition and machine learning (e.g., Farhangfar et al. (2007) and Zhang (2011)). García-Laencina et al. (2010, p. 280) even go as far as saying:

> In classification tasks with missing values, the main objective of an imputation method must be to help to enhance the classification accuracy.

Though this advice may seem logical at first glance, it may actually favor rather strange imputation methods. Let us consider a short example. Suppose that we want to discriminate left-handed people from right-handed people based on their ability to ride a bicycle. Since the relation between these two will be weak, the classification accuracy in complete data will be close to zero. Now suppose that we are missing most of the information on bicycle riding ability, and that we impute the missing information. Depending in our imputation method, we may obtain almost any classification accuracy between zero and 100%. Accuracy will be maximal if we impute missing bicycle-riding ability by right- versus left-handedness. This is, of course, an absurd imputation method, but one that is actually the best method according to García-Laencina et al. (2010), as it provides us with the highest classification accuracy.

The main message is that we cannot evaluate imputation methods by their ability to re-create the true data, or by their ability to optimize classification accuracy. Imputation is not prediction.

2.5.2 Simulation designs and performance measures

The advantageous properties of multiple imputation are only guaranteed if the imputation method used to create the missing data is proper. Equations 2.21–2.23 describe the conditions needed for proper imputation.

Checking the validity of statistical procedures is often done by simulation. There are generally two mechanisms that influence the observed data, the sampling mechanism and the missing data mechanism. Simulation can address sampling mechanism separately, the missing data mechanism separately, and both mechanisms combined. This leads to three general simulation designs.

1. *Sampling mechanism only.* The basic simulation steps are: choose Q, take samples $Y^{(s)}$, perform complete-data analysis, estimate $\hat{Q}^{(s)}$ and $U^{(s)}$ and calculate the outcomes aggregated over s.

2. *Sampling and missing data mechanism combined.* The basic simulation steps are: choose Q, take samples $Y^{(s)}$, generate incomplete data $Y_{obs}^{(s,t)}$, impute, estimate $\hat{Q}^{(s,t)}$ and $T^{(s,t)}$ and calculate outcomes aggregated over s and t.

3. *Missing data mechanism only.* The basic simulation steps are: choose $(\hat{Q}, U)$, generate incomplete data $Y_{obs}^{(t)}$, impute, estimate $(\bar{Q}, \bar{U})^{(t)}$ and $B^{(t)}$ and calculate outcomes aggregated over t.

A popular procedure for testing missing data applications is design 2 with settings $s = 1, \ldots, 1000$ and $t = 1$. As this design does not separate the two mechanisms, any problems found may result from both the sampling and the missing data mechanism. Design 1 does not address the missing data, and is primarily of interest to study whether any problems are attributable to the complete data analysis. Design 3 addresses the missing data mechanism only, and thus allows for a more detailed assessment of any problem caused by the imputation step. An advantage of this procedure is that no population model is needed. Brand et al. (2003) describe this procedure in more detail.

Three outcomes are generally of interest:

1. *Bias*. The bias is calculated as the average difference between the value of the estimand and the value of the estimate. The bias should be close to zero.

2. *Coverage*. The 95% coverage is calculated as the percentage of cases where the value of the estimand is located within the 95% confidence interval around the estimate, called the hit rate. A more efficient alternative is to average the Bayesian probability coverages (Rubin and Schenker, 1986a). The 95% coverage should be 95% or higher. Coverages below 90% are considered undesirable.

3. *Confidence interval length*. The length of the 95% confidence is an indicator of statistical efficiency. The length should be as small as possible, but not so small that coverage falls below 95%.

Section 3.2.3 shows how these quantities can be calculated.

2.6 When to use multiple imputation

Should we always use multiple imputation for the missing data? We probably could, but there are good alternatives in some situations. Section 1.6 already discussed some approaches not covered in this book, each of which has its merits. This section revisits complete case analysis. Apart from being simple to apply, it can be a viable alternative to multiple imputation in particular situations.

Suppose that the complete data model is a regression with outcome Y and predictors X. If the missing data occur in Y only, complete case analysis and multiple imputation are equivalent, so then complete case analysis may be preferred. Multiple imputation gains an advantage over complete case analysis if additional predictors for Y are available that are not part of X.

Under MCAR, complete case analysis is unbiased. It is also efficient if the missing data occur in Y only. Efficiency of complete case analysis declines if X contains missing values, which may result in inflated type II error rates. Complete case analysis can perform quite badly under MAR and some MNAR cases (Schafer and Graham, 2002), but there are two special cases where it can outperform multiple imputation.

The first special case occurs if the probability to be missing does not depend on Y. Under the assumption that the complete data model is correct, the regression coefficients are free of bias (Glynn and Laird, 1986; Little, 1992; King et al., 2001). This holds for any type of regression analysis, and for missing data in both Y and X. Since the missing data rate may depend on X, complete case analysis will in fact work in a relevant class of MNAR models. White and Carlin (2010) confirmed the superiority of complete case analysis by simulation. The differences were often small, and multiple imputation gained the upper hand as more predictive variables were included. The property is useful though in practice.

The second special case holds only if the complete data model is logistic regression. Suppose that the missing data are confined to either a dichotomous Y or to X, but not to both. Assuming that the model is correctly specified, the regression coefficients (except the intercept) are unbiased if only the complete cases are analyzed as long as the probability to be missing depends only on Y and not on X (Vach, 1994). This property provides the statistical basis of the estimation of the odds ratio from case-control studies in epidemiology. If missing data occur in both Y and X the property does not hold.

At a minimum, application of listwise deletion should be a conscious decision of the analyst, and should preferably be accompanied by an explicit

statement that the missing data fit in one of the three categories described above.

2.7 How many imputations?

One of the distinct advantages of multiple imputation is that it can produce unbiased estimates with correct confidence intervals with a low number of imputed datasets, even as low as $m = 2$. Multiple imputation is able to work with low m since it enlarges the between-imputation variance B by a factor $1/m$ before calculating the total variance in $T = \bar{U} + (1 + m^{-1})B$.

The classic advice is to use a low number of imputation, somewhere between 3 and 5 for moderate amounts of missing information. Several authors have recently looked at the influence of m on various aspects of the results. The picture emerging from this work is that it could be beneficial to set m higher, somewhere in the range of 20–100 imputations. This section reviews the relevant work in the area.

The advice for low m rests on the following argument. Multiple imputation is a simulation technique, and hence $\bar{Q}$ and its variance estimate T are subject to simulation error. Setting $m = \infty$ causes all simulation error to disappear, so $T_\infty < T_m$ if $m < \infty$. The question is when T_∞ is close enough to T_m. (Rubin, 1987a, p. 114) showed that the two variances are related by

$$T_m = \left(1 + \frac{\gamma_0}{m}\right) T_\infty \tag{2.38}$$

where γ_0 is the (true) population fraction of missing information. This quantity is equal to the expected fraction of observations missing if Y is a single variable without covariates, and commonly less than this if there are covariates that predict Y. For example, for $\gamma_0 = 0.3$ (e.g., a single variable with 30% missing) and $m = 5$ we find that the calculated variance T_m is $1 + 0.3/5 = 1.06$ times (i.e., 6%) larger than the ideal variance T_∞. The corresponding confidence interval would thus be $\sqrt{1.06} = 1.03$ (i.e., 3%) longer than the ideal confidence interval based on $m = \infty$. Increasing m to 10 or 20 would bring the factor down 1.5% and 0.7%, respectively. The argument is that "the additional resources that would be required to create and store more than a few imputations would not be well spent" (Schafer, 1997, p. 107), and "in most situations there is simply little advantage to producing and analyzing more than a few imputed datasets" (Schafer and Olsen, 1998, p. 549).

Royston (2004) observed that the length of the confidence interval also depends on ν, and thus on m (cf. Equation 2.30). He suggested to base the criterion for m on the confidence coefficient $t_\nu\sqrt{T}$, and proposed that the coefficient of variation of $\ln(t_\nu\sqrt{T})$ should be smaller than 0.05. This effectively

constrains the range of uncertainty about the confidence interval to roughly within 10%. This rule requires m to be "at least 20 and possibly more."

Graham et al. (2007) investigated the effect of m on the statistical power of a test for detecting a small effect size (<0.1). Their advice is to set m high in applications where high statistical power is needed. For example, for $\gamma_0 = 0.3$ and $m = 5$ the statistical power obtained is 73.1% instead of the theoretical value of 78.4%. We need $m = 20$ to increase the power to 78.2%. In order to have an attained power within 1% of the theoretical power, then for fractions of missing information $\gamma = (0.1, 0.3, 0.5, 0.7, 0.9)$ we need to set $m = (20, 20, 40, 100, >100)$, respectively.

Bodner (2008) explored the variability of three quantities under various m: the width of the 95% confidence interval, the p-value, and λ, the proportion of variance attributable to the missing data. Bodner selected m such that the width of the 95% confidence interval is within 10% of its true value 95% of the time. For $\lambda = (0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9)$, he recommends $m = (3, 6, 12, 24, 59, 114, 258)$, respectively. Since the true λ is unknown, Bodner suggested the proportion of complete cases as a conservative estimate of λ.

The starting point of White et al. (2011b) is that all essential quantities in the analysis should be reproducible within some limit, including confidence intervals, p-values and estimates of the fraction of missing information. They take a quote from Von Hippel (2009) as a rule of thumb: *the number of imputations should be similar to the percentage of cases that are incomplete.* This rule applies to fractions of missing information of up to 0.5. If $m \approx 100\lambda$, the following properties will hold for a parameter β:

1. The Monte Carlo error of $\hat{\beta}$ is approximately 10% of its standard error;
2. The Monte Carlo error of the test statistic $\hat{\beta}/\text{se}(\hat{\beta})$ is approximately 0.1;
3. The Monte Carlo error of the p-value is approximately 0.01 when the true p-value is 0.05.

White et al. (2011b) suggest these criteria provide an adequate level of reproducibility in practice.

The idea of reproducibility is sensible, the rule is simple to apply, so there is much to commend it. One potential difficulty might be that the percentage of complete cases is sensitive to the number of variables in the data. If we extend the active dataset by adding more variables, then the percentage of complete cases can only drop. An alternative would be to use the average missing data rate as a less conservative estimate. More experience is needed for how well rules like these work in practice.

Theoretically it is always better to use higher m, but this involves more computation and storage. Setting m high may not be worth the extra wait. Imputing a dataset in practice often involves trial and error to adapt and refine the imputation model. Such initial explorations do not require large m. It is convenient to set $m = 5$ during model building, and increase m only after being satisfied with the model for the "final" round of imputation. So

if calculation is not prohibitive, we may set m to the average percentage of missing data. The substantive conclusions are unlikely to change as a result of raising m beyond $m = 5$.

2.8 Exercises

1. *Nomogram.* Construct a graphic representation of Equation 2.27 that allows the user to convert λ and γ for different values of ν. What influence does ν have on the relation between λ and γ?

2. *Models.* Explain the difference between the response model and the imputation model.

3. *Listwise deletion.* In the `airquality` data, predict `Ozone` from `Wind` and `Temp`. Now randomly delete the half of the wind data above 10 mph, and randomly delete half of the temperature data above 80°F.

   (a) Are the data MCAR, MAR or MNAR?

   (b) Refit the model under listwise deletion. Do you notice a change in the estimates? What happens to the standard errors?

   (c) Would you conclude that listwise deletion provides valid results here?

   (d) If you add a quadratic term to the model, would that alter your conclusion?

4. *Number of imputations.* Consider the `nhanes` dataset in `mice`.

   (a) Use the functions `ccn()` to calculate the number of complete cases. What percentage of the cases is incomplete?

   (b) Impute the data with `mice` using the defaults with `seed=1`, predict `bmi` from `age`, `hyp` and `chl` by the normal linear regression model, and pool the results. What are the proportions of variance due to the missing data for each parameter? Which parameters appear to be most affected by the nonresponse?

   (c) Repeat the analysis for `seed=2` and `seed=3`. Do the conclusions remain the same?

   (d) Repeat the analysis with $m = 50$ with the same seeds. Would you prefer this analysis over those with $m = 5$? Explain why.

5. *Number of imputations (continued).* Continue with the data from the previous exercise.

(a) Write an `R` function that automates the calculations of the previous exercise. Let `seed` run from 1 to 100 and let `m` take on values `m=c(3,5,10,20,30,40,50,100, 200)`.

(b) Plot the estimated proportions of explained variance due to missing data for the `age`-parameter against m. Based on this graph, how many imputations would you advise?

(c) Check White's conditions 1 and 2 (cf. Section 2.7). For which m do these conditions true?

(d) Does this also hold for categorical data? Use the `nhanes2` to study this.

6. *Automated choice of m.* Write an `R` function that implements the methods discussed in Section 2.7.

Chapter 3

Univariate missing data

Chapter 2 described the theory of multiple imputation. This chapter looks into ways of creating the actual imputations. In order to avoid unnecessary complexities at this point, the text is restricted to univariate missing data. The incomplete variable is called the *target* variable. Thus, in this chapter there is only one variable with missing values. The consequences of the missing data depend on the role of the target variables within the complete data model that is applied to the imputed data.

There are many ways to create imputations, but only a few of those lead to valid statistical inferences. This chapter outlines ways to check the correctness of a procedure, and how this works out for selected procedures. Most of the methods are designed to work under the assumption that the relations within the missing parts are similar to those in the observed parts, or more technically, the assumption of ignorability. The chapter closes with a description of some alternatives of what we might do when that assumption is suspect.

3.1 How to generate multiple imputations

This section illustrates five ways to create imputations for a single incomplete continuous target variable. We use dataset number 88 in Hand et al. (1994), which is also part of the `MASS` library under the name `whiteside`. Mr. Whiteside of the UK Building Research Station recorded the weekly gas consumption (in 1000 cubic feet) and average external temperature (in °C) at his own house in south-east England for two heating seasons (1960 and 1961). The house thermostat was set at 20°C throughout.

Figure 3.1a plots the observed data. More gas is needed in colder weeks, so there is an obvious relation in the data. The dataset is complete, but for the sake of argument suppose that the gas consumption in row 47 of the data is missing. The temperature at this deleted observation is equal to 5°C. How would we create multiple imputations for the missing gas consumption?

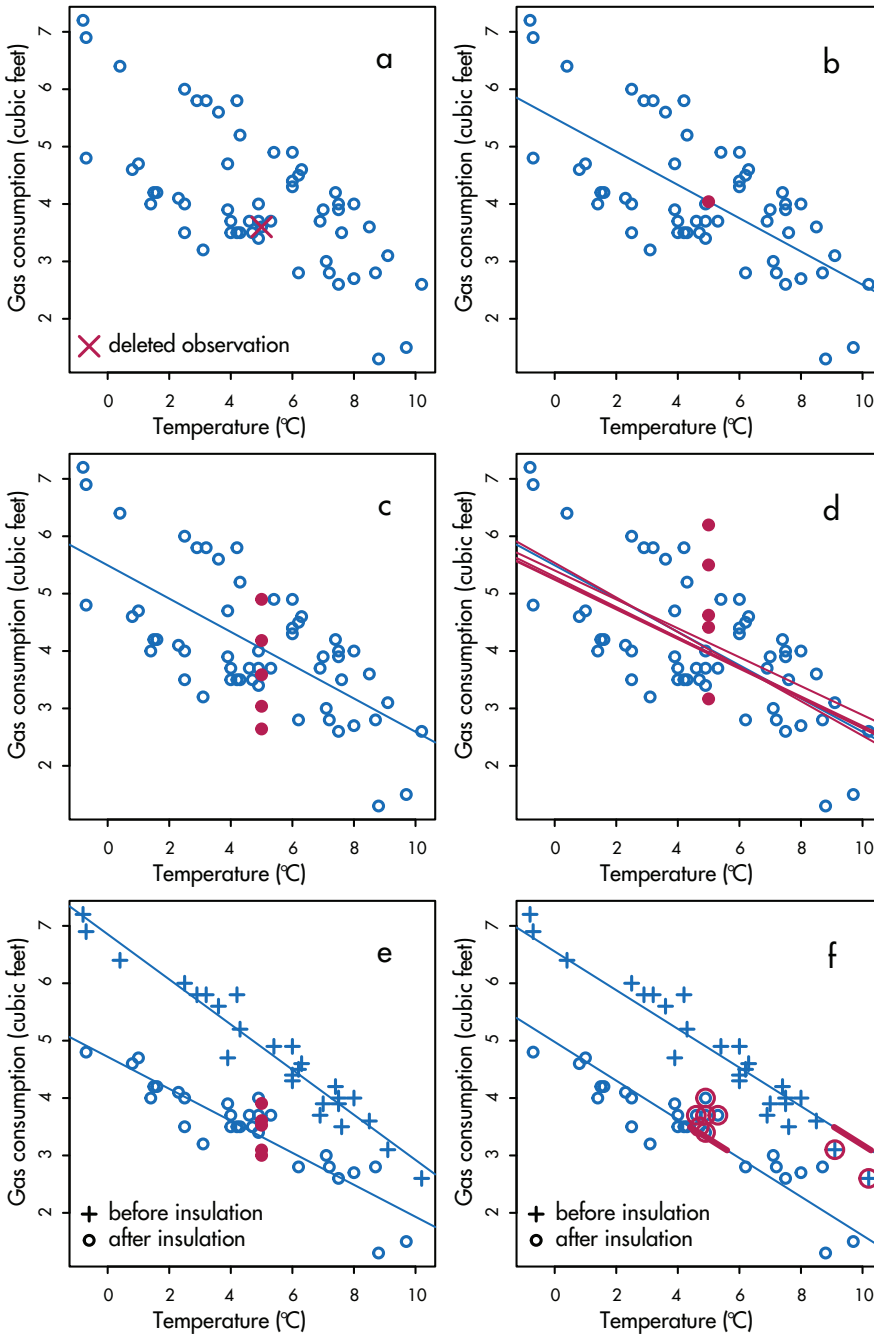


Figure 3.1: Five ways to impute missing gas consumption for a temperature of 5°C: (a) no imputation; (b) predict; (c) predict + noise; (d) predict + noise + parameter uncertainty; (e) two predictors; (f) drawing from observed data.

3.1.1 Predict method

A first possibility is to calculate the regression line, and take the imputation from the regression line. The estimated regression line is equal to $y = 5.49 - 0.29x$, so the value at $x = 5$ is $5.49 - 0.29 \times 5 = 4.04$. Figure 3.1b shows where the imputed value is. This is actually the "best" value in the sense that it is the most likely one under the regression model. However, even the best value may differ from the actual (unknown) value. In fact, we are uncertain about the true gas consumption. Predicted values, however, do not portray this uncertainty, and therefore cannot be used as multiple imputations.

3.1.2 Predict + noise method

We can improve upon the prediction method by adding an appropriate amount of random noise to the predicted value. Let us assume that the observed data are normally distributed around the regression line. The estimated standard deviation in the Whiteside data is equal to 0.86 cubic feet. The idea is now to draw a random value from a normal distribution with a mean of zero and a standard deviation of 0.86, and add this value to the predicted value. The underlying assumption is that the distribution of gas consumption of the incomplete observation is identical to that in the complete cases.

We can repeat the draws to get multiple synthetic values around the regression line. Figure 3.1c illustrates five such drawn values. On average, the synthetic values will be equal to the predicted value. The variability in the values reflects that fact that we cannot accurately predict gas consumption from temperature.

3.1.3 Predict + noise + parameter uncertainty

Adding noise is a major step forward, but not quite right. The method in the previous section requires that the intercept, the slope and the standard deviation of the residuals are known. However, the values of these parameters are typically unknown, and hence must be estimated from the data. If we had drawn a different sample from the same population, then our estimates for the intercept, slope and standard deviation would be different, perhaps slightly. The amount of extra variability is strongly related to the sample size, with smaller samples yielding more variable estimates.

The parameter uncertainty also needs to be included in the imputations. There are two main methods for doing so. Bayesian methods draw the parameters directly from their posterior distributions, whereas bootstrap methods resample the observed data and re-estimate the parameters from the resampled data.

Figure 3.1d shows five sampled regression lines calculated by the Bayesian method. Imputed values are now defined as the predicted value of the sampled line added with noise, as in Section 3.1.2.

3.1.4 A second predictor

The dataset actually contains a second predictor that indicates whether the house was insulated or not. Incorporating this extra information reduces the uncertainty of the imputed values.

Figure 3.1e shows the same data, but now flagged according to insulation status. Two regression lines are shown, one for the insulated houses and the other for the non-insulated houses. It is clear that less gas is needed after insulation. Suppose we know that the external temperature is 5°C *and* that the house was insulated. How do we create multiple imputation given these two predictors?

We apply the same method as in Section 3.1.3, but now using the regression line for the insulated houses. Figure 3.1e shows the five values drawn for this method. As expected, the distribution of the imputed gas consumption has shifted downward. Moreover, its variability is lower, reflecting that fact that gas consumption can be predicted more accurately as insulation status is also known.

3.1.5 Drawing from the observed data

Figure 3.1f illustrates an alternative method to create imputations. As before, we calculate the predicted value at 5°C for an insulated house, but now select a small number of candidate donors from the observed data. The selection is done such that the predicted values are close. We then randomly select one donor from the candidates, and use the *observed* gas consumption that belongs to that donor as the synthetic value. The figure illustrates the candidate donors, not the imputations.

This method is known as *predictive mean matching*, and always finds values that have been actually observed in the data. The underlying assumption is that within the group of candidate donors gas consumption has the same distribution in donors and receivers. The variability between the imputations over repeated draws is again a reflection of the uncertainty of the actual value.

3.1.6 Conclusion

In summary, prediction methods are not suitable to create multiple imputations. Both the inherent prediction error and the parameter uncertainty should be incorporated into the imputations. Adding a relevant extra predictor reduces the amount of uncertainty, and leads to more efficient estimates later on. The text also highlights an alternative that draws imputations from the observed data. The imputation methods discussed in this chapter are all variations on this basic idea.

3.2 Imputation under the normal linear normal

3.2.1 Overview

For univariate Y we write lowercase y for Y. Any predictors in the imputation model are collected in X. Symbol X_{obs} indicates the subset of n_1 rows of X for which y is observed, and X_{mis} is the complementing subset of n_0 rows of X for which y is missing. The vector containing the n_1 observed data in y is denoted by y_{obs}, and the vector of n_0 imputed values in y is indicated by $\dot{y}$. This section reviews four different ways of creating imputations under the normal linear model. The four methods are:

1. *Predict.* $\dot{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are least squares estimates calculated from the observed data. Section 1.3.4 named this regression imputation. In `mice` this method is available as `"norm.predict"`.

2. *Predict + noise.* $\dot{y} = \hat{\beta}_0 + X_{\text{mis}}\hat{\beta}_1 + \dot{\epsilon}$, where $\dot{\epsilon}$ is randomly drawn from the normal distribution as $\dot{\epsilon} \sim N(0, \hat{\sigma}^2)$. Section 1.3.5 named this stochastic regression imputation. In `mice` this method is available as `"norm.nob"`.

3. *Bayesian multiple imputation.* $\dot{y} = \dot{\beta}_0 + X_{\text{mis}}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ and $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are random draws from their posterior distribution, given the data. Section 3.1.3 named this "predict + noise + parameters uncertainty." The method is available as `"norm"`.

4. *Bootstrap multiple imputation.* $\dot{y} = \dot{\beta}_0 + X_{\text{mis}}\dot{\beta}_1 + \dot{\epsilon}$, where $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$, and where $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ are the least squares estimates calculated from a bootstrap sample taken from the observed data. This is an alternative way to implement "predict + noise + parameters uncertainty." The method is available as `"norm.boot"`.

3.2.2 Algorithms ♠

The calculations of the first two methods are straightforward and do not need further explanation. This section describes the algorithms used to introduce sampling variability into the parameters estimates of the imputation model.

The Bayesian method draws $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}$ from their respective posterior distributions. Box and Tiao (1973, section 2.7) explains the Bayesian theory behind the normal linear model. We use the method that draws imputations under the normal linear model using the standard non-informative priors for each of the parameters. Given these priors, the required inputs are:

- y_{obs}, the $n_1 \times 1$ vector of observed data in the incomplete (or target) variable y;

Algorithm 3.1: Bayesian imputation under the normal linear model.♠

1. Calculate the cross-product matrix $S = X'_{\text{obs}} X_{\text{obs}}$.
2. Calculate $V = (S + \text{diag}(S)\kappa)^{-1}$, with some small κ.
3. Calculate regression weights $\hat{\beta} = V X'_{\text{obs}} y_{\text{obs}}$.
4. Draw a random variable $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$.
5. Calculate $\dot{\sigma}^2 = (y_{\text{obs}} - X_{\text{obs}}\hat{\beta})'(y_{\text{obs}} - X_{\text{obs}}\hat{\beta})/\dot{g}$.
6. Draw q independent $N(0,1)$ variates in vector $\dot{z}_1$.
7. Calculate $V^{1/2}$ by Cholesky decomposition.
8. Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z}_1 V^{1/2}$.
9. Draw n_0 independent $N(0,1)$ variates in vector $\dot{z}_2$.
10. Calculate the n_0 values $\dot{y} = X_{\text{mis}}\dot{\beta} + \dot{z}_2\dot{\sigma}$.

- X_{obs}, the $n_1 \times q$ matrix of predictors of rows with observed data in y;
- X_{mis}, the $n_0 \times q$ matrix of predictors of rows with missing data in y.

The algorithm assumes that both X_{obs} and X_{mis} contain no missing values. Chapter 4 deals with the case where X_{obs} and X_{mis} also could be incomplete.

Algorithm 3.1 is adapted from Rubin (1987a, p. 167), and is implemented in the function `mice.impute.norm()` of the `mice` package. Any drawn values are identified with a dot above the symbol, so $\dot{\beta}$ is a value of β drawn from the posterior distribution. The algorithm uses a ridge parameter κ to evade problems with singular matrices. This number should be set to a positive number close to zero, e.g., $\kappa = 0.0001$. For some data, larger κ may be needed. High values of κ, e.g., $\kappa = 0.1$, may introduce a systematic bias toward the null, and should thus be avoided.

The bootstrap is a general method for estimating sampling variability through resampling the data (Efron and Tibshirani, 1993). Algorithm 3.2 calculates univariate imputations by drawing a bootstrap sample from the complete part of the data, and subsequently takes the least squares estimates given the bootstrap sample as a "draw" that incorporates sampling variability into the parameters (Heitjan and Little, 1991). Compared to the Bayesian method, the bootstrap method avoids the Choleski decomposition and does not need to draw from the χ^2-distribution.

Algorithm 3.2: Imputation under the normal linear model with bootstrap.♠

1. Draw a bootstrap sample $(\dot{y}_{\text{obs}}, \dot{X}_{\text{obs}})$ of size n_1 from $(y_{\text{obs}}, X_{\text{obs}})$.
2. Calculate the cross-product matrix $\dot{S} = \dot{X}'_{\text{obs}}\dot{X}_{\text{obs}}$.
3. Calculate $\dot{V} = (\dot{S} + \text{diag}(\dot{S})\kappa)^{-1}$, with some small κ.
4. Calculate regression weights $\dot{\beta} = \dot{V}\dot{X}'_{\text{obs}}\dot{y}_{\text{obs}}$.
5. Calculate $\dot{\sigma}^2 = (\dot{y}_{\text{obs}} - \dot{X}_{\text{obs}}\dot{\beta})'(\dot{y}_{\text{obs}} - \dot{X}_{\text{obs}}\dot{\beta})/(n_1 - q - 1)$.
6. Draw n_0 independent $N(0,1)$ variates in vector $\dot{z}_2$.
7. Calculate the n_0 values $\dot{y} = X_{\text{mis}}\dot{\beta} + \dot{z}_2\dot{\sigma}$.

3.2.3 Performance

Which of these four imputation methods of Section 3.2 is best? In order to find out we conduct a small simulation experiment in which we calculate the bias, coverage and average confidence interval width for each method. We keep close to the original data by assuming that $\beta_0 = 5.49$, $\beta_1 = -0.29$ and $\sigma = 0.86$ are the population values. These values are used to generate artificial data with known properties.

In `R` we create a small function `createdata()` that randomly draws artificial data from a linear model with given parameters:

```
> ### create data
> createdata <- function(beta0=5.49, beta1=-0.29, sigma=0.86,
                  n=50, mx=5, sdx=3) {
    x <- round(rnorm(n,mean=mx,sd=sdx),1)
    eps <- rnorm(n,mean=0,sd=sigma)
    y <- round(beta0 + x * beta1 + eps, 1)
    return(data.frame(x=x, y=y))
  }
```

The values `mx` and `sdx` were chosen to mimic the variation in temperature in the `whiteside` data. Rounding is applied to get close to the actual data, but is not strictly needed.

The next step is to create missing data. Here we use a simple random missing data mechanism (MCAR) to create approximately 50% of missing random data. The percentage of missing data is set high to get a clear picture of the properties of each method.

```
> ### make 50% random missing data
> makemissing <- function(data, p=0.5){
```

```
  rx <- rbinom(nrow(data), 1, p)
  data[rx==0,"y"] <- NA
  return(data)
}
```

A next piece is a small test function that calls `mice()` and calculates various statistics of interest. We concentrate on the slope parameter β_1 of the regression line, but we could have equally opted for β_0 or σ. We specify `ridge=0` to set the ridge parameter κ to zero, and thus get an unbiased estimate, and set the default number of imputations to $m = 5$.

```
> ### test function for three imputation functions
> test.impute <- function(data, m=5, method="norm", ...){
    imp <- mice(data, method=method, m=m, print=FALSE,
         ridge=0, ...)
    fit <- with(imp, lm(y~x))
    est <- pool(fit)
    tab <- summary(est)
    return(tab["x",c("est","se","lo 95","hi 95","fmi",
        "lambda")])
  }
```

The following function puts everything together:

```
> simulate <- function(nsim=10, seed=41872){
    set.seed(seed)
    res <- array(NA,dim=c(4, nsim, 6))
    dimnames(res) <- list(c("predict","pred + noise",
                    "Bayes MI","boot MI"),
                         as.character(1:10000),
                         c("est","se","lo 95","hi 95","fmi",
                              "lambda"))
    im <- c("norm.predict","norm.nob","norm","norm.boot")
    for(i in 1:nsim){
      data <- createdata()
      data <- makemissing(data)
      res[1,i,] <- test.impute(data, method=im[1], m=2)
      res[2,i,] <- test.impute(data, method=im[2])
      res[3,i,] <- test.impute(data, method=im[3])
      res[4,i,] <- test.impute(data, method=im[4])
    }
    return(res)
  }
```

Performing 10000 simulations is now done by calling `simulate()`, thus

```
> res <- simulate(10000)
```

Some postprocessing of the results is needed to calculate the statistics of interest. The means of the six outcomes per method are calculated as

```
> apply(res,c(1,3),mean,na.rm=TRUE)

                 est     se  lo 95  hi 95    fmi lambda
predict      -0.289 0.0285 -0.346 -0.232 0.0407  0.000
pred + noise -0.289 0.0522 -0.402 -0.176 0.4086  0.343
Bayes MI     -0.289 0.0646 -0.446 -0.133 0.5694  0.499
boot MI      -0.289 0.0621 -0.442 -0.136 0.5636  0.494
```

This code uses the `apply()` function on the three-dimensional array `res` that contains the simulation results. The bias of β_1 per method is calculated as the average deviation from the true value $\beta_1 = -0.29$ by

```
> # bias of beta1
> true <- -0.29
> bias <- rowMeans(res[,,1] - true)
> bias
     predict pred + noise     Bayes MI      boot MI
    0.000809     0.000607     0.000462     0.001043
```

The coverage is calculated as the percentage that the true β_1 is within the 95% confidence interval by

```
> isin <- res[,,3] < true & true < res[,,4]
> cov <- rowMeans(isin)
> cov
     predict pred + noise     Bayes MI      boot MI
       0.654        0.907        0.951        0.944
```

The average width of the 95% confidence interval for β_1 is calculated by

```
> intwidth <- res[,,4] - res[,,3]
> aiw <- rowMeans(intwidth)
> aiw
     predict pred + noise     Bayes MI      boot MI
       0.115        0.226        0.313        0.306
```

Table 3.1 summarizes the results. The predict method produces an unbiased estimate, but its confidence interval is much too short, leading to substantial undercoverage and p-values that are "too significant." This result clearly illustrates the problems already noted in Section 2.5.1. The "predict + noise" method performs better, but the coverage of 0.907 is too low. However, both the Bayesian and bootstap methods are correct, with very few differences between them. Also, the λ parameter estimated by both is close to 0.50, which is the correct result.

Table 3.1: Missing y—properties of β_1 under imputation of missing y by four methods for the normal linear model ($n_{\text{sim}} = 10000$)

Method	Bias	Coverage	CI Width	$\hat{\gamma}$	$\hat{\lambda}$
Predict	.001	.654	.115	.041	.000
Predict+noise	.001	.907	.226	.409	.343
Bayesian MI	.000	.951	.313	.569	.499
Bootstrap MI	.001	.944	.306	.564	.494
CCA	.001	.951	.250		

Table 3.2: Missing x—properties of β_1 under imputation of missing x by four methods for the normal linear model ($n_{\text{sim}} = 10000$)

Method	Bias	Coverage	CI Width	$\hat{\gamma}$	$\hat{\lambda}$
Predict	−.101	.361	.161	.041	.000
Predict+noise	.001	.928	.202	.287	.230
Bayesian MI	.008	.955	.254	.459	.391
Bootstrap MI	−.001	.948	.238	.410	.345
CCA	−.001	.951	.250		

For comparison, the table also include the results of complete case analysis (CCA). Complete case analysis is a correct analysis here (Little and Rubin, 2002). In fact, it is the most efficient choice for this problem as it yields the shortest confidence interval (cf. Section 2.6). This result does not hold more generally, however. In realistic situations involving more covariates multiple imputation will rapidly catch up and pass complete case analysis.

While the "predict" method is simple and fast, the variance estimate is too low. Various methods have been proposed to correct the variance (Lee et al., 1994; Fay, 1996; Rao, 1996; Schafer and Schenker, 2000). Though such methods require special adaptation of formulas to calculate the variance, they may be useful when the missing data are restricted to the outcome.

It is straightforward to adapt the simulations to other, perhaps more interesting situations. For example, we could investigate the effect of missing data in the explanatory x instead of the outcome variable by changing the function `makemissing()` and rerun the simulations. Table 3.2 displays the results. The "predict" method is now severely biased, whereas the other methods remain unbiased. The confidence interval of "predict+noise" is still too short, but less than in Table 3.1. The Bayesian and bootstrap methods are both correct, in the sense that they are unbiased and have appropriate coverage. It seems that the bootstrap method is slightly more efficient, but a direct comparison is difficult because the coverages differ.

We could increase the number of explanatory variables and the number of imputations m to see how much the average confidence interval width would

shrink. It is also easy to apply more interesting missing data mechanisms, such as those discussed in Section 3.2.4. Data can be generated from skewed distributions, the sample size n can be varied and so on. Much simulation work is available (Rubin and Schenker, 1986b; Rubin, 1987a).

3.2.4 Generating MAR missing data

Just making random missing data is not always interesting. We obtain more informative simulations if the missingness probability is a function of the observed, and possibly of the unobserved, information. This section considers some methods for creating univariate missing data. These could form building blocks for missing data generation in a multivariate context by combining them. See Van Buuren et al. (2006, appendix B) for details.

Let us first consider three methods to create missing data in artificial data. The data are generated as 1000 draws from the bivariate normal distribution $P(Y_1, Y_2)$ with means $\mu_1 = \mu_2 = 5$, variances $\sigma_1^2 = \sigma_2^2 = 1$, and covariance $\sigma_{12} = 0.6$. We assume that all values generated are positive. Missing data in Y_2 can be created in many ways. Let R_2 be the response indicator for Y_2. We study three examples, each of which affects the distribution in different ways:

$$\text{MARRIGHT} \quad : \quad \text{logit}(\Pr(R_2 = 0)) = -5 + Y_1 \tag{3.1}$$

$$\text{MARMID} \quad : \quad \text{logit}(\Pr(R_2 = 0)) = 0.75 - |Y_1 - 5| \tag{3.2}$$

$$\text{MARTAIL} \quad : \quad \text{logit}(\Pr(R_2 = 0)) = -0.75 + |Y_1 - 5| \tag{3.3}$$

where $\text{logit}(p) = \log(p) - \log(1 - p)$ with $0 \leq p \leq 1$ is the logit function. Its inverse $\text{logit}^{-1}(x) = \exp(x)/(1 + \exp(x))$ is known as the logistic function.

Generating missing data under these models in `R` can be done in three steps: calculate the missingness probability of each data point, make a random draw from the binomial distribution, and set the corresponding observations to `NA`. The following script creates missing data according to MARRIGHT:

```
> logistic <- function(x) exp(x)/(1+exp(x))
> set.seed(32881)
> n <- 10000
> y <- mvrnorm(n=n,mu=c(5,5),Sigma=matrix(c(1,0.6,0.6,1),
        nrow=2))
> p2.marright <- 1 - logistic(-5 + y[,1])
> r2.marright <- rbinom(n, 1, p2.marright)
> yobs <- y
> yobs[r2.marright==0, 2] <- NA
```

Figure 3.2 displays the probability of being missing under the three MAR mechanisms. All mechanisms yield approximately 50% of missing data, but do so in very different ways.

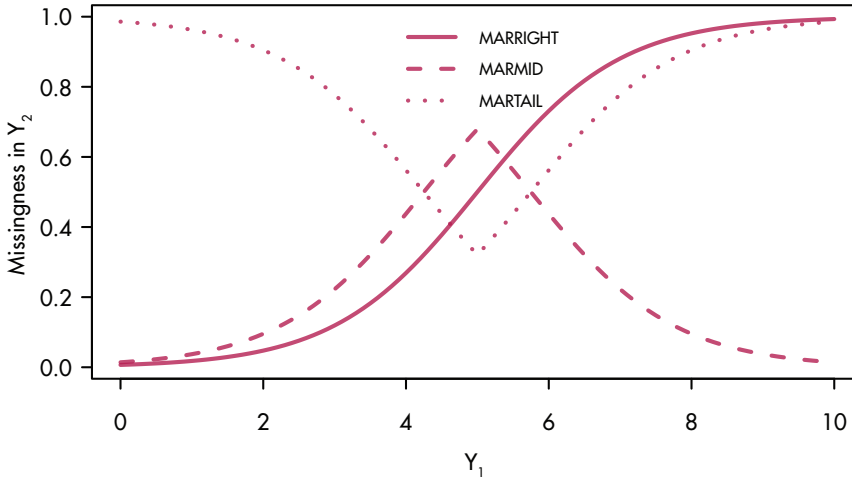


Figure 3.2: Probability that Y_2 is missing as a function of the values of Y_1 under three models for the missing data.

Figure 3.3 displays the distributions of Y_2 under the three models. MARRIGHT deletes more high values, so the distribution of the observed data shifts to the left. MARMID deletes more data in the center, so the variance of the observed data grows, but the mean is not affected. MARTAIL shows the reverse effect. The variance of observed data reduces because of the missing data.

These mechanisms are more extreme than we are likely to see in practice. Not only is there a strong relation between Y_1 and R_2, but the percentage of missing data is also quite high (50%). On the other hand, if methods perform well under these data deletion schemes, they will also do so in less extreme situations that are more likely to be encountered in practice.

The objective of Exercise 3.1 is to study the behavior of the four imputation methods under these missing data mechanisms.

3.2.5 Conclusion

Based on Tables 3.1 and 3.2 both the "predict" and "predict + noise" methods fail in terms of underestimating the variance. If the missing data occur in y only, then it is possible to correct the variance formulas of the "predict" method. However, when the missing data occur in X, the "predict" method is severely biased, so then variance correction is not useful. The two methods that account for the uncertainty of the imputation model provide statistically correct inferences. For missing y, the efficiency of these methods is less than theoretically possible, presumably due to simulation error. If the

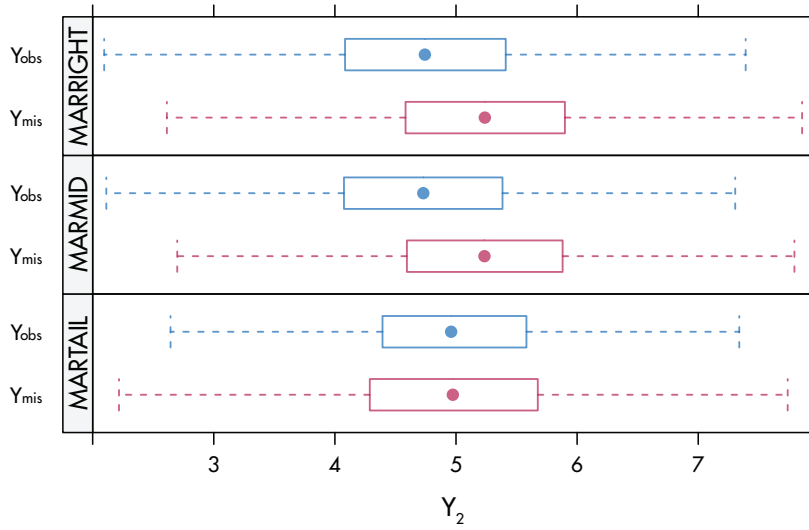


Figure 3.3: Box plot of Y_2 separated for the observed and missing parts under three models for the missing data based on $n = 10000$.

missing data occur in X, multiple imputation may be more efficient than complete case analysis (see also Section 2.6).

It is always better to include parameter uncertainty, either by the Bayesian or the bootstrap method. The effect of doing so will diminish with increasing sample size (Exercise 2), so for estimates based on a large sample one may opt for the simpler "predict + noise" method. Note that in subgroup analyses, the large-sample requirement applies to the subgroup size, and not to the total sample size.

3.3 Imputation under non-normal distributions

3.3.1 Overview

The imputation methods discussed in Section 3.2 produce imputations drawn from a normal distribution. In practice the data could be skewed, long tailed, non-negative, bimodal or rounded, to name some deviations from normality. This creates an obvious mismatch between observed and imputed data which could adversely affect the estimates of interest.

The effect of non-normality is generally small for measures that rely on

the center of the distribution, like means or regression weights, but it could be substantial for estimates like a variance or a percentile. In general, normal imputations appear to be robust against violations of normality. Demirtas et al. (2008) found that flatness of the density, heavy tails, non-zero peakedness, skewness and multimodality do not appear to hamper the good performance of multiple imputation for the mean structure in samples $n > 400$, even for high percentages (75%) of missing data in one variable. The variance parameter is more critical though, and could be off-target in smaller samples.

A sensible approach is to transform the data toward normality before imputation, and back-transform them after imputation. A beneficial side effect of transformation is that the relation between x and y may become closer to a linear relation. Sometimes applying a simple function to the data, like the logarithmic or inverse transform, is all that is needed. More generally, the transformation could be made to depend on known covariates like age and sex, for example as done in the LMS model (Cole and Green, 1992) or the GAMLSS model (Rigby and Stasinopoulos, 2005).

It is also possible to directly draw imputations from non-normal distributions. Liu (1995) proposed methods for drawing imputations under the t-distribution instead of the normal. He and Raghunathan (2006) created imputations by drawing from Tukey's gh-distribution, which can take many shapes. Demirtas and Hedeker (2008a) investigated the behavior of methods for drawing imputation from the Beta and Weibull distributions. Likewise, Demirtas and Hedeker (2008b) took draws from Fleishman polynomials, which allows for combinations of left and right skewness with platykurtic and leptokurtic distributions.

3.3.2 Imputation from the t-distribution ♠

Though it is not standard functionality, some methods for non-normal data can be used in `mice` by calling the `gamlss()` function. The `gamlss` package (Stasinopoulos and Rigby, 2007) contains over 60 built-in distributions. Each of these comes with a function to draw random variates. One may construct a new univariate imputation function that calls `gamlss()` to do the fitting and the imputation.

We illustrate this by creating a function with a residual t-distribution instead of the normal. The t-distribution is favored for more robust statistical modeling in a variety of settings (Lange et al., 1989). Extending to other distributions than the t-distribution and building more complicated `gamlss` imputation models can be done along the same lines.

```
> mice.impute.TF <- function(y, ry, x,
                     gamlss.trace = FALSE, ...)
  {
    require(gamlss)
```

```
    # prepare data
    xobs <- x[ry, , drop = FALSE]
    xmis <- x[!ry, , drop = FALSE]
    yobs <- y[ry]
    n1 <- sum(ry)
    n0 <- sum(!ry)

    # draw bootstrap sample
    s <- sample(n1, n1, replace = TRUE)
    dotxobs <- xobs[s, , drop = FALSE]
    dotyobs <- yobs[s]
    dotxy <- data.frame(dotxobs, y = dotyobs)

    # fit the gamlss model
    fit <- gamlss(y ~ ., data = dotxy, family = TF,
                  trace = gamlss.trace, ...)
    yhat <- predict(fit, data=dotxy, newdata = xmis)
    sigma <- exp(coef(fit, "sigma"))
    nu <- exp(coef(fit, "nu"))

    # draw the imputations
    return(rTF(n0, yhat, sigma, nu))
  }
>
```

3.3.3 Example ♠

Van Buuren and Fredriks (2001) observed unexplained kurtosis in the distribution of head circumference in children. Rigby and Stasinopoulos (2006) fitted a t-distribution to these data, and observed a substantial improvement of the fit.

Figure 3.4 plots the data for Dutch boys aged 1–2 years. Due to the presence of several outliers, the t distribution with 6.7 degrees of freedom fits the data substantially better than the normal distribution (Akaike Information Criterion (AIC): 2974.5 (normal model) versus 2904.3 (t-distribution). If the outliers are genuine data, then the t-distribution should provide imputations that are more realistic than the normal.

We create a synthetic dataset by imputing head circumference of the same 755 boys. Imputation is easily done with the following steps: append the data with a duplicate, create missing data in `hc` and run `mice()` calling our newly programmed `TF` method as follows:

```
> library(gamlss)
> data(db)
> data <- subset(db, age>1&age<2, c("age","head"))
```

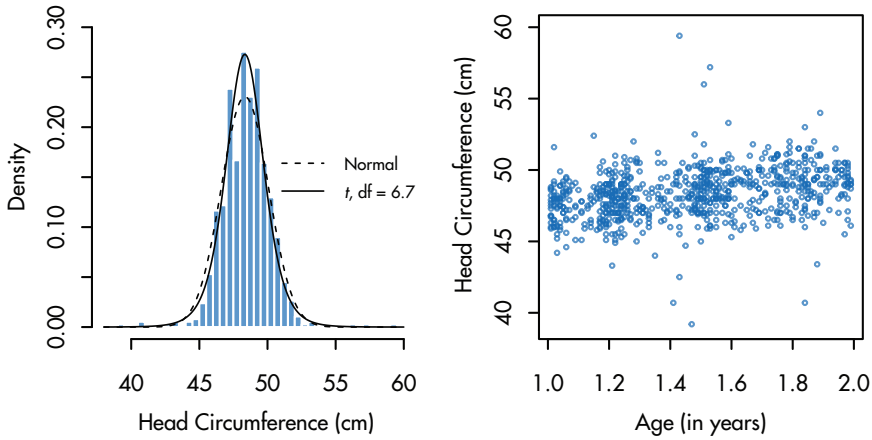


Figure 3.4: Measured head circumference of 755 Dutch boys aged 1–2 years (Fredriks et al., 2000a).

```
> names(data) <- c("age","hc")
> synthetic <- rep(c(FALSE,TRUE), each=nrow(data))
> data2 <- rbind(data, data)
> data2[synthetic,"hc"] <- NA
> imp <- mice(data2, m=1, meth="TF", seed=36650, print=FALSE)
> syn <- subset(complete(imp), synthetic)
```

Figure 3.5 is the equivalent of Figure 3.4, but now calculated from the synthetic data. Both configurations are similar. As expected, some outliers also occur in the imputed data. There are some small differences. The estimated degrees of freedom varies over replications, and hovers around the value of 6.7 estimated from the observed data. For this replication, it is somewhat lower (4.4). Moreover, it appears that the distribution of the imputed data is slightly more "well behaved" than in the observed data. The typical rounding patterns seen in the real measurements are not present in the imputed data. Though these are small differences, they may be of relevance in particular analyses.

3.4 Predictive mean matching

3.4.1 Overview

Predictive mean matching subsamples from the observed data. The method calculates the predicted value of target variable Y according to the specified

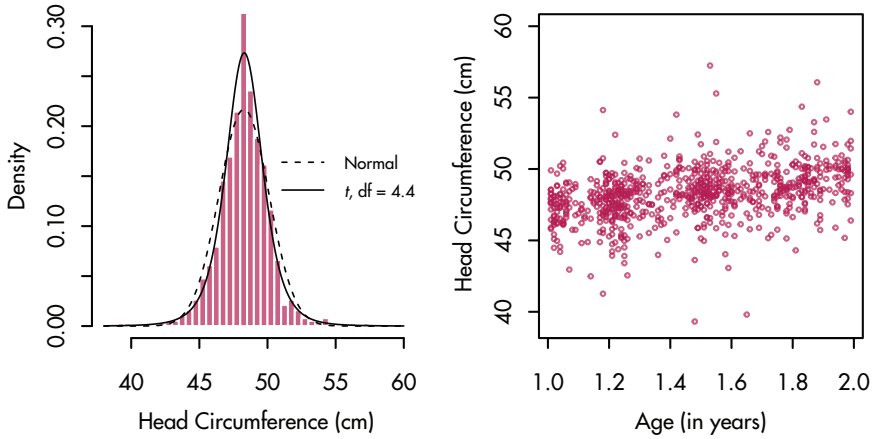


Figure 3.5: Fully synthetic data of head circumference of 755 Dutch boys aged 1–2 years using a t-distribution.

imputation model. For each missing entry, the method forms a small set of candidate donors (typically with 1, 3 or 10 members) from all complete cases that have predicted values close to the predicted value for the missing entry. A random draw is made among the candidates, and the observed value of the donors is taken to replace the missing value. The assumption made is that within each set, the receivers' data follows the same distribution as the candidates' data.

Predictive mean matching is an easy-to-use and versatile method. It is fairly robust to transformations of the target variable, so imputing $\log(Y)$ often yields results similar to imputing $\exp(Y)$. The method also allows for discrete target variables. Imputations are based on values observed elsewhere, so they are realistic. Imputations outside the observed data range will not occur, thus evading problems with meaningless imputations (e.g., negative body height). The model is implicit (Little and Rubin, 2002), which means that there is no need to define an explicit model for the distribution of the missing values. Because of this, predictive mean matching is less vulnerable to model misspecification than the methods discussed in Sections 3.2 and 3.3.

Figure 3.6 illustrates the robustness of predictive mean matching relative to the normal model. The figure displays the body mass index (BMI) of children aged 0–2 years. BMI rapidly increases during the first half year of life, has a peak around 1 year and then slowly drops at ages when the children start to walk. The imputation model is, however, incorrectly specified, being linear in age. Imputations created under the normal model display in an incorrect slowly rising pattern, and contain several implausible values. In contrast, the imputations created by predictive mean matching follow the data quite nicely, even though the predictive mean itself is clearly off-target for some of the

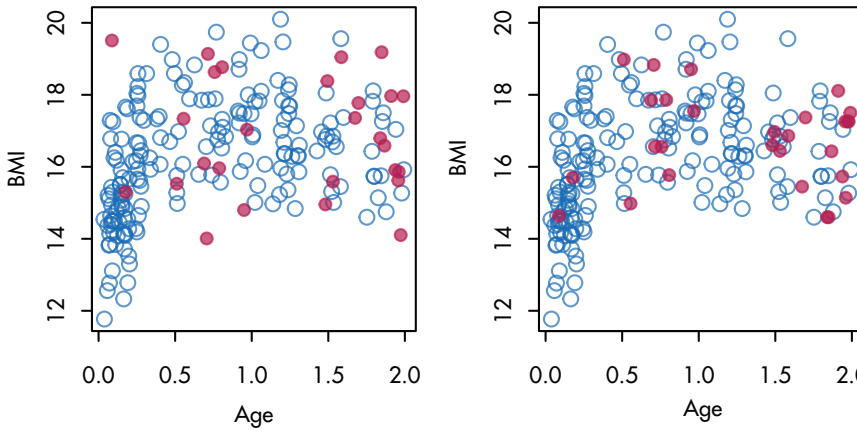


Figure 3.6: Robustness of predictive mean matching (right) relative to imputation under the linear normal model (left).

ages. This example shows that predictive mean matching is robust against misspecification, where the normal model is not.

Predictive mean matching is an example of a hot deck method, where values are imputed using values from the complete cases matched with respect to some metric. The expression "hot deck" literally refers to a pack of computer control cards containing the data of the cases that are in some sense close. Reviews of hot deck methods can be found in Ford (1983), Brick and Kalton (1996), Koller-Meinfelder (2009), Andridge and Little (2010) and De Waal et al. (2011, pp. 249–255, 349–355).

3.4.2 Computational details ♠

Various metrics are possible to define the distance between the cases. The predictive mean matching metric was proposed by Rubin (1986) and Little (1988). It is particularly useful for missing data applications because it is optimized for each target variable separately. The predicted value is generally a convenient one-number summary of the important information that relates to the target. Calculation is straightforward, and it is easy to include nominal and ordinal variables.

Once the metric has been defined, there are various ways to select the donor. Let $\hat{y}_i$ denote the predicted value of the rows with an observed y_i where $i = 1, \ldots, n_1$. Likewise, let $\hat{y}_j$ denote the predicted value of the rows with missing y_j where $j = 1, \ldots, n_0$. Andridge and Little (2010) distinguish four methods:

1. Choose a threshold η, and take all i for which $|\hat{y}_i - \hat{y}_j| < \eta$ as candidate

donors for imputing j. Randomly sample one donor from the candidates, and take its y_i as replacement value.

2. Take the closest candidate, i.e., the case i for which $|\hat{y}_i - \hat{y}_j|$ is minimal as the donor. This is known as "nearest neighbor hot deck," "deterministic hot deck" or "closest predictor."

3. Find the d candidates for which $|\hat{y}_i - \hat{y}_j|$ is minimal, and sample one of them. Usual values for d are 3, 5 and 10. There is also an adaptive method to specify the number of donors (Schenker and Taylor, 1996).

4. Sample one donor with a probability that depends on $|\hat{y}_i - \hat{y}_j|$ (Siddique and Belin, 2008).

The obvious danger of predictive mean matching is the duplication of the same donor value many times. This problem is more likely to occur if the sample is small, or if there are many more missing data than observed data in a particular region of the predicted value. Such unbalanced regions are more likely if the proportion of incomplete cases is high, or if the imputation model contains variables that are very strongly related to the missingness.

Some simulation work is available about the different strategies for defining the set of candidate donors. Setting $d = 1$ is generally considered to be too low, as it may reselect the same donor over and over again. Predictive mean matching performs very badly when d is small and there are lots of ties for the predictors among the individuals to be imputed. The reason is that the tied individuals all get the same imputed value in each imputed dataset when $d = 1$ (Ian White, personal communicaton). Setting d to a high value (say $n/10$) alleviates the duplication problem, but may introduce bias since the likelihood of bad matches increases. Schenker and Taylor (1996) evaluated $d = 3$, $d = 10$ and an adaptive scheme. The adaptive method was slightly better than using a fixed number of candidates, but the differences were small. The authors note that there may also be situations where adaptive estimation could be more beneficial, but further work on this issue is still lacking.

Another issue is that the traditional method does not work for a small number of predictors. Heitjan and Little (1991) report that for just two predictors the results were "disastrous." The problem has received little attention over the years. The cause of the problem appears to be related to the type of matching used. More precisely, it is useful to distinguish three types of matching:

1. *Type 0*: $\hat{y} = X_{\text{obs}}\hat{\beta}$ is matched to $\hat{y}_j = X_{\text{mis}}\hat{\beta}$;

2. *Type 1*: $\hat{y} = X_{\text{obs}}\hat{\beta}$ is matched to $\dot{y}_j = X_{\text{mis}}\dot{\beta}$;

3. *Type 2*: $\dot{y} = X_{\text{obs}}\dot{\beta}$ is matched to $\dot{y}_j = X_{\text{mis}}\dot{\beta}$.

Here $\hat{\beta}$ is the estimate of β, while $\dot{\beta}$ is a value randomly drawn from the posterior distribution of β. Type 0 matching ignores the sampling variability

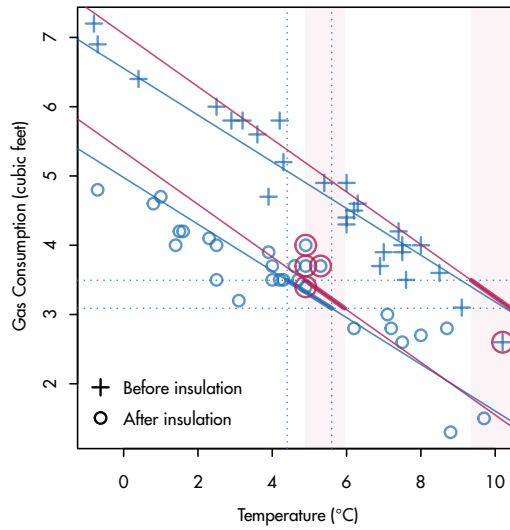


Figure 3.7: Selection of candidate donors in predictive mean matching with the stochastic matching distance.

in $\hat{\beta}$, leading to improper imputations. Type 2 matching appears to solve this. However, it is insensitive to the process of taking random draws of β if there are only a few variables. In the extreme case, with a single X, the set of candidate donors based on $|\dot{y}_i - \dot{y}_j|$ remains unchanged under different values of $\dot{\beta}$, so the same donor(s) get selected too often. Type 1 matching is a small but nifty adaptation of the matching distance that seems to alleviate the problem. The difference with Type 0 and Type 2 matching is that in Type 1 matching only $X_{\text{mis}}\dot{\beta}$ varies stochastically and does not cancel out anymore. As a result $\dot{\eta}$ incorporates between-imputation variation. In retrospect, it is interesting to note that Type 1 matching was in fact already described by Little (1988, eq. 4). Evidently it disappeared from the literature, only to reappear two decades later in the works of Koller-Meinfelder (2009, p. 43) and White et al. (2011b, p. 383).

Figure 3.7 is a graphic illustration of the method using the `whiteside` data. The predictor is equal to 5°C and we use $\eta = 0.6$. The thick blue line indicates the area of the target variable where matches should be sought. The blue parts of the figure are considered fixed. The red lines correspond to random draws that incorporate the sampling uncertainty. The two light red "bands" indicate the area where matches are permitted. In this particular instance, five candidate donors are found, four from the subgroup "After insulation" and one from the subgroup "Before insulation." The last step is to make a random draw among these five candidates. The red parts in the figure move

Algorithm 3.3: Predictive mean matching with a Bayesian β and a stochastic matching distance (Type 1 matching).♠

1. Calculate $\dot{\beta}$ and $\hat{\beta}+$ by Steps 1-8 of Algorithm 3.1.
2. Calculate $\dot{\eta}(i,j) = |X_{\text{obs},[i]}\hat{\beta} - X_{\text{mis},[j]}\dot{\beta}|$ with $i = 1,\ldots,n_1$ and $j = 1,\ldots,n_0$.
3. Construct n_0 sets Z_j, each containing d candidate donors, from Y_{obs} such that $\sum_d \dot{\eta}(i,j)$ is minimum for all $j = 1,\ldots,n_0$. Break ties randomly.
4. Draw one donor i_j from Z_j randomly for $j = 1,\ldots,n_0$.
5. Calculate imputations $\dot{y}_j = y_{i_j}$ for $j = 1,\ldots,n_0$.

over different imputed datasets, and thus the set of candidates also varies over the imputations.

The data point at coordinate (10.2, 2.6) is one of the candidate donors. This point differs from the incomplete unit in both temperature and insulation status, yet it is selected as a candidate donor. The advantage of including the point is that closer matches in terms of the predicted values are possible. As long as the distribution of the target in different bands can be considered similar, including points from different bands might be beneficial. It is not known how robust the method is against violations of this assumption. There is a considerable literature on matching methods in the context of observational studies and file matching where this problem is recognized (Rubin, 2006; D'Orazio et al., 2006). It is possible to rule out certain combinations. We could also contemplate techniques for constrained matching that prohibit candidates from being used more than once as a donor. These methods are largely unknown for missing data applications, but might well merit further exploration in practical applications.

3.4.3 Algorithm ♠

Algorithm 3.3 provides the steps used in predictive mean matching using Bayesian parameter draws for β. We can create the bootstrap version of this algorithm that will also evade the need to draw β along the same lines as Algorithm 3.2. Given that the number of candidate donors and the model for the mean is provided by the user, the algorithm does not need an explicit specification of the distribution, and is fully automatic.

Table 3.3 repeats the simulation experiment done in Tables 3.1 and 3.2 for predictive mean matching for three different choices of the number d of candidate donors. The results are as good as we can get, and are essentially

Table 3.3: Properties of β_1 under multiple imputation by predictive mean matching and $m = 5$ ($n_{\text{sim}} = 10000$)

Method	Bias	Coverage	CI Width	$\hat{\gamma}$	$\hat{\lambda}$
Missing y					
PMM $d = 1$	$-.001$	.950	.309	.566	.496
PMM $d = 3$	$-.001$	.951	.313	.571	.501
PMM $d = 10$	$-.001$	.950	.313	.570	.500
Missing x					
PMM $d = 1$	.006	.957	.253	.459	.391
PMM $d = 3$	.007	.954	.253	.459	.391
PMM $d = 10$	.007	.954	.250	.454	.386
CCA	$-.001$	.951	.250		

equivalent to the results obtained for the theoretically superior Bayesian and bootstrap methods discussed earlier. The number of candidate donors does not affect performance in this simulation, which is in line with the results of Schenker and Taylor (1996).

3.4.4 Conclusion

Predictive mean matching with $d = 3$ is the default in `mice()` for continuous data. The method is robust against misspecification of the imputation model, yet performs as well as theoretically superior methods. In the context of missing covariate data, Marshall et al. (2010a) concluded that predictive mean matching "produced the least biased estimates and better model performance measures." Another simulation study that addressed skewed data concluded that "MICE-PMM may be the preferred MI approach provided that less than 50% of the cases have missing data and the missing data are not MNAR" (Marshall et al., 2010b). The method works best with large samples, and provides imputations that possess many characteristics of the complete data. Predictive mean matching cannot be used to extrapolate beyond the range of the data, or to interpolate within the range of the data if the data at the interior are sparse. Also, it may not perform well with small datasets. Bearing these points in mind, predictive mean matching is a great all-around method with exceptional properties.

3.5 Categorical data

3.5.1 Overview

Imputation of missing categorical data is possible under the broad class of generalized linear models (McCullagh and Nelder, 1989). For incomplete binary variables we use *logistic regression*, where the outcome probability is modeled as

$$\Pr(y_i = 1|X_i, \beta) = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)} \tag{3.4}$$

A categorical variable with K unordered categories is imputed under the *multinomial logit model*

$$\Pr(y_i = k|X_i, \beta) = \frac{\exp(X_i\beta_k)}{\sum_{k=1}^{K} \exp(X_i\beta_k)} \tag{3.5}$$

for $k = 1, \ldots, K$, where β_k varies over the categories and where $\beta_1 = 0$ to identify the model. A categorical variable with K ordered categories is imputed by the *ordered logit model*, or *proportional odds model*

$$\Pr(y_i \leq k|X_i, \beta, \tau_k) = \frac{\exp(\tau_k - X_i\beta)}{1 + \exp(\tau_k - X_i\beta)} \tag{3.6}$$

with the slope β is identical across categories, but the intercepts τ_k differ. For identification, we set $\tau_1 = 0$. The probability of observing category k is written as

$$\Pr(y_i = k|X_i) = \Pr(y_i \leq k|X_i) - \Pr(y_i \leq k - 1|X_i) \tag{3.7}$$

where the model parameters β, τ_k and τ_{k-1} are suppressed for clarity. Scott Long's book provides an introduction to these methods. The practical application of these techniques in `R` is treated in Aitkin et al. (2009).

The general idea is to estimate the probability model on the subset of the observed data, and draw synthetic data according to the fitted probabilities to impute the missing data. The parameters are typically estimated by iteratively reweighted least squares. As before, the variability of the model parameters β and $\tau_2, \ldots, \tau_K$ introduces additional uncertainty that needs to be incorporated into the imputations.

Algorithm 3.4 provides the steps for an approximate Bayesian imputation method using logistic regression. The method assumes that the parameter vector β follows a multivariate normal distribution. Although this is true in large samples, the distribution can in fact be far from normal for modest n_1, for large q or for predicted probabilities close to 0 or 1. The procedure is also approximate in the sense that it does not draw the estimated covariance V matrix. It is possible to define an explicit Bayesian method for

Algorithm 3.4: Imputation of a binary variable by means of Bayesian logistic regression.♠

1. Estimate regression weights $\hat{\beta}$ from $(y_{\text{obs}}, X_{\text{obs}})$ by iteratively reweighted least squares.
2. Obtain V, the unscaled estimated covariance matrix of $\hat{\beta}$.
3. Draw q independent $N(0,1)$ variates in vector $\dot{z}_1$.
4. Calculate $V^{1/2}$ by Cholesky decomposition.
5. Calculate $\dot{\beta} = \hat{\beta} + \dot{z}_1 V^{1/2}$.
6. Calculate n_0 predicted probabilities $\dot{p} = 1/(1 + \exp(-X_{\text{mis}}\dot{\beta}))$.
7. Draw n_0 random variates from the uniform distribution $U(0,1)$ in the vector u.
8. Calculate imputations $\dot{y}_j = 1$ if $u_j \leq \dot{p}_j$, and $\dot{y}_j = 0$ otherwise, where $j = 1, \ldots, n_0$.

drawing β and V from their exact posteriors. This method is theoretically preferable, but as it requires more elaborate modeling, it does not easily extend to other regression situations. In `mice` the algorithm is implemented as function `mice.impute.logreg()`, and is used as the default for binary data.

It is easy to construct a bootstrap version that avoids some of the difficulties in Algorithm 3.4. Prior to estimating $\hat{\beta}$, we include a step that draws a bootstrap sample from Y_{obs} and X_{obs}. Steps 2–5 can then be replaced by equating $\dot{\beta} = \hat{\beta}$.

The algorithms for imputation of variables with more than two categories follow the same structure. In `mice` the multinomial logit model is estimated by the `multinom()` function in the `nnet` package. The ordered logit model is estimated by the `polr()` function of the `MASS` package. Even though the ordered model uses fewer parameters, it is often more difficult to estimate. In cases where `polr()` fails to converge, `multinom()` will take over its duties. See Venables and Ripley (2002) for more details on both functions.

3.5.2 Perfect prediction ♠

There is a long-standing technical problem in models with categorical outcomes, known as *separation* or *perfect prediction* (Albert and Anderson, 1984; Lesaffre and Albert, 1989). The standard work by Hosmer and Lemeshow (2000, pp. 138–141) discussed the problem, but provided no solution. The problem occurs, for example, when predicting the presence of a disease from

Table 3.4: Artificial data demonstrating complete separation. Adapted from White et al. (2010).

Disease	Symptom	
	Yes	No
Yes	100	100
No	0	100
Unknown	100	100

a set of symptoms. If one of the symptoms (or a combination of symptoms) always leads to the disease, then we can perfectly predict the disease for any patient who has the symptom(s).

Table 3.4 contains an artificial numerical example. Having the symptom always implies the disease, so knowing that the patient has the symptom will allow perfect prediction of the disease status. When such data are analyzed, most software will print out a warning message and produce unusually large standard errors.

Now suppose that in a new group of 200 patients (100 in each symptom group) we know only the symptom and impute disease status. Under MAR, we should impute all 100 cases with the symptom to the diseased group, and divide the 100 cases without the symptom randomly over the diseased and non-diseased groups. However, this is not what happens in Algorithm 3.4. The estimate of V will be very large as a result of separation. If we naively use this V then $\dot{\beta}$ in step 5 effectively covers both positive and negative values equally likely. This results in either correctly 100 imputations in `Yes` or incorrectly 100 imputations in `No`, thereby resulting in bias in the disease probability.

The problem has recently gained attention. There are at least six different approaches to perfect prediction:

1. Eliminate the variable that causes perfect prediction.
2. Take $\hat{\beta}$ instead of $\dot{\beta}$.
3. Use penalized regression with Jeffreys prior in step 2 of Algorithm 3.4 (Firth, 1993; Heinze and Schemper, 2002).
4. Use the bootstrap, and then apply method 1.
5. Use data augmentation, a method that concatenates pseudo-observations with a small weight to the data, effectively prohibiting infinite estimates (Clogg et al., 1991; White et al., 2010).
6. Apply the explicit Bayesian method with a suitable weak prior. Gelman et al. (2008) recommend using independent Cauchy distributions on all logistic regression coefficients.

Eliminating the most predictive variable is generally undesirable in the

context of imputation, and may in fact bias the relation of interest. Option 2 does not yield proper imputations, and is therefore not recommended. Option 3 provides finite estimates, but has been criticized as not being well interpretable in a regression context (Gelman et al., 2008) and computationally inefficient (White et al., 2010). Option 4 corrects method 1, and is simple to implement. Options 5 and 6 have been recommended by White et al. (2010) and Gelman et al. (2008), respectively.

Methods 4, 5 and 6 all solve a major difficulty in the construction of automatic imputation techniques. It is not yet clear whether one of these methods is superior. Version 2.0 of `mice` implemented option 5.

3.6 Other data types

3.6.1 Count data

Examples of count data include the number of children in a family or the number of crimes committed. The minimum value is zero. Imputing incomplete count data should produce non-negative synthetic replacement values. Count data can be imputed in various ways:

1. Predictive mean matching (cf. Section 3.4)
2. Ordered categorical imputation (cf. Section 3.5)
3. (Zero-inflated) Poisson regression (Raghunathan et al., 2001)
4. (Zero-inflated) negative binomial regression (Royston, 2009)

Poisson regression is a class of models that is widely applied in biostatistics. The Poisson model can be thought of as the sum of the outcomes from a series of independent flips of the coin. The negative binomial is a more flexible model that is often applied an as alternative to account for overdispersion. Zero-inflated versions of both models can be used if the number of zero values is larger than expected. See Scott Long (1997) for an accessible description of these models and Aitkin et al. (2009) for computational aspects. The models are special cases of the generalized linear model, and do not bring new issues compared to, say, logistic regression imputation. It is straightforward to adapt the computer code in Section 3.3.2 to create an imputation function for Poisson regression and the negative binomial model using the `gamlss` package.

It not yet clear whether one of the four methods consistently outperforms the others. Until more detailed results become available, my advice is to use predictive mean matching for count data.

3.6.2 Semi-continuous data

Semi-continuous data have a high mass at one point (often zero) and a continuous distribution over the remaining values. An example is the number of cigarettes smoked per day, which has a high mass at zero because of the non-smokers, and an often highly skewed unimodal distribution for the smokers. The difference with count data is gradual. Semi-continuous data are typically treated as continuous data, whereas count data are generally considered discrete.

Imputation of semi-continuous variables needs to reproduce both the point mass and continuously varying part of the data. One possibility is to apply a general-purpose method that preserves distributional features, like predictive mean matching (cf. Section 3.4).

An alternative is to model the data in two parts. The first step is to determine whether the imputed value is zero or not. The second step is only done for those with a nonzero value, and consists of drawing a value from the continuous part. Olsen and Schafer (2001) developed an imputation technique by combining a logistic model for the discrete part, and a normal model for the continuous part, possibly after a normalizing transformation. A more general two-part model was developed by Javaras and Van Dyk (2003), who extended the standard general location model (Olkin and Tate, 1961) to impute partially observed semi-continuous data.

Yu et al. (2007) evaluated nine different procedures. They found that predictive mean matching performs well, provided that a sufficient number of data points in the neighborhood of the incomplete data are available. So care is required for small samples and/or rare events. It is not known whether predictive mean matching outperforms two-part models in general.

3.6.3 Censored, truncated and rounded data

An observation y_i is censored if its value is only partly known. In *right-censored* data we only know that $y_i > a_i$ for a censoring point a_i. In *left-censored* data we only know that $y_i \leq b_i$ for some known censoring point b_i, and in *interval censoring* we know $a_i \leq y_i \leq b_i$. Right-censored data arise when the true value is beyond the maximum scale value, for example, when body weight exceeds the scale maximum, say 150 kg. When y_i is interpreted as time taken to some event (e.g., death), right-censored data occur when the observation period ends before the event has taken place. Left and right censoring may cause floor and ceiling effects. Rounding data to fewer decimal places results in interval-censored data.

Truncation is related to censoring, but differs from it in the sense that value below (left truncation) or above (right truncation) the truncation point is not recorded at all. For example, if persons with a weight in excess of 150 kg are removed from the sample, we speak of truncation. The fact that observations are entirely missing turns the truncation problem into a missing

data problem. Truncated data are less informative than censored data, and consequently truncation has a larger potential to distort the inferences of interest.

The usual approach for dealing with missing values in censored and truncated data is to delete the incomplete records, i.e., complete case analysis. In the event that time is the censored variable, consider the following two problems:

- *Censored event times.* What would have been the uncensored event time if no censoring had taken place?
- *Missing event times.* What would have been the event time and the censoring status if these had been observed?

The problem of censored event times has been studied extensively. There are many statistical methods that can analyze left- or right-censored data directly, collectively known as *survival analysis*. Kleinbaum and Klein (2005), Hosmer et al. (2008) and Allison (2010) provide useful introductions into the field. Survival analysis is the method of choice if censoring is restricted to the single outcomes. The approach is, however, less suited for censored predictors or for multiple interdependent censored outcomes. Van Wouwe et al. (2009) discuss an empirical example of such a problem. The authors are interested in knowing time interval between resuming contraception and cessation of lactation in young mothers who gave birth in the last 6 months. As the sample was cross-sectional, both contraception and lactation were subject to censoring. Imputation could be used to impute the hypothetically uncensored event times in both durations, and this allowed a study of the association between the uncensored event times.

The problem of missing event times is relevant if the event time is unobserved. The censoring status is typically also unknown if the event time is missing. Missing event times may be due to happenstance, for example, resulting from a technical failure of the instrument that measures event times. Alternatively, the missing data could have been caused by truncation, where all event times beyond the truncation point are set to missing. It will be clear that the optimal way to deal with the missing events data depends on the reasons for the missingness. Analysis of the complete cases will systematically distort the analysis of the event times if the data are truncated.

Imputation of right-censored data has received most attention to date. In general, the method aims to find new (longer) event times that would have been observed had the data not been censored. Let n_1 denote the number of observed failure times, let $n_0 = n - n_1$ denote the number of censored event times and let $t_1, \ldots, t_n$ be the ordered set of failure and censored times. For some time point t, the *risk set* $R(t) = t_i > t$ for $i = 1, \ldots, n$ is the set of event and censored times that is longer than t. Taylor et al. (2002) proposed two imputation strategies for right-censored data:

1. *Risk set imputation.* For a given censored value t construct the risk set

$R(t)$, and randomly draw one case from this set. Both the failure time and censoring status from the selected case are used to impute the data.

2. *Kaplan–Meier imputation.* For a given censored value t construct the risk set $R(t)$ and estimate the Kaplan–Meier curve from this set. A randomly drawn failure time from the Kaplan–Meier curve is used for imputation.

Both methods are asymptotically equivalent to the Kaplan–Meier estimator after multiple imputation with large m. The adequacy of imputation procedures will depend on the availability of possible donor observations, which diminishes in the tails of the survival distribution. The Kaplan–Meier method has the advantage that nearly all censored observations are replaced by imputed failure times. In principle, both Bayesian and bootstrap methods can be used to incorporate model uncertainty, but in practice only the bootstrap has been used.

Hsu et al. (2006) extended both methods to include covariates. The authors fitted a proportional hazards model and calculated a risk score as a linear combination of the covariates. The key adaptation is to restrict the risk set to those cases that have a risk score that is similar to the risk score of censored case, an idea similar to predictive mean matching. A donor group size with $d = 10$ was found to perform well, and Kaplan–Meier imputation was superior to risk set imputation across a wide range of situations.

Algorithm 3.5 is based on the KIMB method proposed by Hsu et al. (2006). The method assumes that censoring status is known, and aims to impute plausible event times for censored observations. Hsu et al. (2006) actually suggested fitting two proportional hazards models, one with survival time as outcome and one with censoring status as outcome, but in order to keep in line with the rest of this chapter, here we only fit the model for survival time. The way in which predictive mean matching is done differs slightly from Hsu et al. (2006).

The literature on imputation methods for censored and rounded data is rapidly evolving. Alternative methods for right-censored data have also been proposed (Wei and Tanner, 1991; Geskus, 2001; Lam et al., 2005; Liu et al., 2011). Lyles et al. (2001), Lynn (2001) and Hopke et al. (2001) concentrated on left-censored data. Imputation of interval-censored data (rounded data) has been discussed quite extensively (Heitjan and Rubin, 1990; Dorey et al., 1993; James and Tanner, 1995; Pan, 2000; Bebchuk and Betensky, 2000; Glynn and Rosner, 2004; Hsu, 2007; Royston, 2007; Chen and Sun, 2010). Imputation of double-censored data, where both the initial and the final times are interval censored, is treated by Pan (2001) and Zhang et al. (2009). By comparison, very few methods have been developed to deal with truncation. Methods for imputing a missing censoring indicator have been proposed by Subramanian (2009, 2011) and Wang and Dinse (2010).

Algorithm 3.5: Imputation of right-censored data using predictive mean matching, Kaplan–Meier estimation and the bootstrap.♠

1. Estimate $\hat{\beta}$ by a proportional hazards model of y given X, where $y = (t, \phi)$ consists of time t and censoring indicator ϕ ($\phi_i = 0$ if t_i is censored).
2. Draw a bootstrap sample $(\dot{y}, \dot{X})$ of size n from (y, X).
3. Estimate $\dot{\beta}$ by a proportional hazards model of $\dot{y}$ given $\dot{X}$.
4. Calculate $\dot{\eta}(i, j) = |X_{[i]}\hat{\beta} - X_{[j]}\dot{\beta}|$ with $i = 1, \ldots, n$ and $j = 1, \ldots, n_0$, where $[j]$ indexes the cases with censored times.
5. Construct n_0 sets Z_j, each containing d candidate donors such that $t_i > t_j$ and $\sum_d \dot{\eta}(i, j)$ is minimum for each $j = 1, \ldots, n_0$. Break ties randomly.
6. For each Z_j, estimate the Kaplan–Meier curve $\hat{S}_j(t)$.
7. Draw n_0 uniform random variates u_j, and take $\dot{t}_j$ from the estimated cumulative distribution function $1 - \hat{S}_j(t)$ at u_j for $j = 1, \ldots, n_0$.
8. Set $\phi_j = 0$ if $\dot{t}_j = t_n$ and $\phi_{t_n} = 0$, else set $\phi_j = 1$.

3.7 Classification and regression trees

3.7.1 Overview

Classification and regression trees (CART) (Breiman et al., 1984) are a popular class of machine learning algorithms. CART models seek predictors and cut points in the predictors that are used to split the sample. The cut points divide the sample into more homogeneous subsamples. The splitting process is repeated on both subsamples, so that a series of splits defines a binary tree. The target variable can be discrete (classification tree) or continuous (regression tree).

Figure 3.8 illustrates a simple CART solution for the `airquality` data. The left-hand figure contains the optimal binary tree for predicting gas consumption from temperature and insulation status. The right-hand side shows the scatterplot in which the five groups are labeled by their terminal nodes.

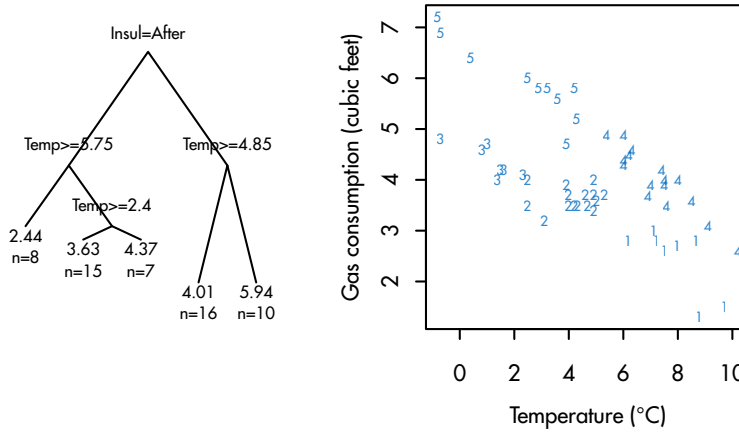


Figure 3.8: Regression tree for predicting gas consumption. The left-hand plot displays the binary tree, whereas the right-hand plot identifies the groups at each end leaf in the data.

3.7.2 Imputation using CART models

CART methods have properties that make them attractive for imputation: they are robust against outliers, can deal with multicollinearity and skewed distributions, and are flexible enough to fit interactions and nonlinear relations. Furthermore, many aspects of model fitting have been automated, so there is "little tuning needed by the imputer" (Burgette and Reiter, 2010).

The idea of using CART methods for imputation has been suggested by a wide variety of authors in a variety of ways. See Saar-Tsechansky and Provost (2007) for an introductory overview. Some investigators (He, 2006; Vateekul and Sarinnapakorn, 2009) simply fill in the mean or mode. The majority of tree-based imputation methods use some form of single imputation based on prediction (Bárcena and Tusell, 2000; Conversano and Cappelli, 2003; Siciliano et al., 2006; Creel and Krotki, 2006; Ishwaran et al., 2008; Conversano and Siciliano, 2009). Multiple imputation methods have been developed by Harrell (2001), who combined it with optimal scaling of the input variables, by Reiter (2005b) and by Burgette and Reiter (2010). Wallace et al. (2010) present a multiple imputation method that averages the imputations to produce a single tree and that does not pool the variances. Parker (2010) investigates multiple imputation methods for various unsupervised and supervised learning algorithms.

To fix ideas, consider how imputations can be created using the tree in Figure 3.8. For a given temperature and insulation status, traverse the tree and find the appropriate terminal node. Form the donor group of all observed cases at the terminal node, randomly draw a case from the donor group, and take its reported gas consumption as the imputed value.

Algorithm 3.6: Imputation under a tree model using the bootstrap.♠

1. Draw a bootstrap sample $(\dot{y}_{\text{obs}}, \dot{X}_{\text{obs}})$ of size n_1 from $(y_{\text{obs}}, X_{\text{obs}})$.
2. Fit $\dot{y}_{\text{obs}}$ by $\dot{X}_{\text{obs}}$ by a tree model $f(X)$.
3. Predict the n_0 terminal nodes g_j from $f(X_{\text{mis}})$.
4. Construct n_0 sets Z_j of all cases at node g_j, each containing d_j candidate donors.
5. Draw one donor i_j from Z_j randomly for $j = 1, \ldots, n_0$.
6. Calculate imputations $\dot{y}_j = y_{i_j}$ for $j = 1, \ldots, n_0$.

The idea is identical to predictive mean matching (cf. Section 3.4), where the "predictive mean" is now calculated by a tree model instead of a regression model. As before, the parameter uncertainty can be incorporated by fitting the tree on a bootstrapped sample.

Algorithm 3.6 describes the major steps of an algorithm for creating imputations using a classification or regression tree. There is considerable freedom at step 2, where the tree model is fitted to the training data $(\dot{y}_{\text{obs}}, \dot{X}_{\text{obs}})$. It may be useful to fit the tree such that the number of cases at each node is equal to some preset number, say 5 or 10. The composition of the donor groups will vary over different bootstrap replications, which incorporates the sampling uncertainty about the tree.

The studies done to date have concentrated on predictive accuracy, which is not a useful criterion in the context of imputation (cf. Section 2.5.1). None of the studies reported coverage statistics. The potential of tree-based methods and other machine learning techniques (Hastie et al., 2009) for creating proper multiple imputations has yet to be explored.

3.8 Multilevel data

3.8.1 Overview

The problem of missing values in the outcome variable in multilevel data has received considerable attention. The multilevel model is actually "made to solve" the problem of incomplete outcomes. There is an extensive literature which often concentrates on the longitudinal case (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Daniels and Hogan, 2008). For more details, see the encyclopedic overview in Fitzmaurice et al. (2009). Most work

concentrates on deriving valid estimates of the parameters of the model, and is generally less concerned with deriving imputations under the multilevel model.

Multiple imputation of multilevel data is a problem that has not yet been fully solved. The imputation literature up to the year 2008 is summarized in Van Buuren (2011). Zhao and Yucel (2009) discuss methods for various types of outcomes, where Andridge (2011) studies the impact of imputation for cluster randomized trials.

Van Buuren (2011) studied the properties of a full Bayesian method for creating imputations under heteroscedastic error variance, and compared it to three ad hoc alternatives: analysis of the complete cases, imputation ignoring the clustering structure and imputation using fixed effects for classes. The Bayesian method recovers the intra-class correlation quite well, even for severe MAR cases and high amounts of missing data in the outcome or the predictor. Though the technique considerably improves upon standard practice, it has been found that it fails to achieve nominal coverage for the fixed effects, especially for small class sizes. More work is needed to solve these issues.

3.8.2 Two formulations of the linear multilevel model ♠

Let the data be divided in K classes or levels, and let y_c denote the n_c vector containing outcomes on units i ($i = 1, \ldots, n_c$) within class c ($c = 1, \ldots, K$). The univariate *linear mixed effects model* (Laird and Ware, 1982) is written as

$$y_c = X_c\beta + Z_c u_c + \epsilon_c \tag{3.8}$$

where X_c is a known $n_c \times p$ design matrix in class c associated with the common $p \times 1$ fixed effects vector β, and where Z_c is a known $n_c \times q$ design matrix in class c associated with the $q \times 1$ random effect vectors u_c. The random effects u_c are independently and interchangeably normally distributed as $u_c \sim N(0, \Omega)$. The number of random effects q is typically smaller than the number of fixed effects p. Symbol ϵ_c denotes the $n_c \times 1$ vector of residuals, which are independently normally distributed as $\epsilon_c \sim N(0, \sigma_c^2 I(n_c))$ for $c = 1, \ldots, K$. It is often assumed that the residual variance is equal for all classes: $\sigma_c^2 = \sigma^2$. In addition, ϵ_c and u_c are assumed to be independent. Equation 3.8 separates the fixed and random effects.

One may also conceptualize the mixed effects model in Equation 3.8 as a two-level model. To see how this works, write the two-level linear model as

$$y_c = Z_c\beta_c + \epsilon_c \qquad \text{level-1} \quad \text{equation} \tag{3.9}$$

where β_c is a $q \times 1$ vector of regression coefficients that vary between the K classes. At level 2, we model β_c by the linear regression model

$$\beta_c = W_c\beta + u_c \qquad \text{level-2} \quad \text{equation} \tag{3.10}$$

where W_c is a $q \times p$ matrix of a special structure (Van Buuren, 2011, p. 175), and where u_c can be interpreted as the $q \times 1$ vector of level 2 residuals.

Equations 3.9 and 3.10 are collectively called the *slopes-as-outcome model* (Bryk and Raudenbush, 1992).

Note that the regression coefficient β is identical in all level-2 classes. Substituting Equation 3.10 into Equation 3.9 yields

$$y_c = Z_c W_c \beta + Z_c u_c + \epsilon_c \tag{3.11}$$

which is a special case of the linear mixed model (Equation 3.8) with $X_c = Z_c W_c$.

Equation 3.8 separates the fixed and random effects, but the same covariates may appear in both X_c and Z_c. This complicates imputation of those covariates. To make matters more complex, X_c can also contain interactions between covariates at level 1 and level 2. In contrast, the slopes-as-outcomes model distinguishes the level 1 from the level 2 predictors. There is no overlap in data between W_c and Z_c. This is the more convenient parameterization in a missing data context.

Missing data can occur in the outcome variable y_c, the level-1 predictors Z_c, the level-2 predictors W_c and the class variable c. Here we concentrate on missing data in y_c or Z_c, but not both. Multivariate missing data are treated elsewhere (Van Buuren, 2011).

3.8.3 Computation ♠

We discuss two methods to impute multilevel data, a Bayesian method and a bootstrap method. The Bayesian method first draws parameters randomly from their appropriate posteriors distributions, and conditional on these draws generates synthetic values. The idea is the same as in Algorithm 3.1.

Suppose that y_{obs} represents the observed outcome data, and let X_{obs} and Z_{obs} be the fixed and random predictors for the complete cases, respectively. Our model allows that the residual variances σ_c^2 vary over the classes. The following parameters must be simulated from the data: $\dot{\beta}$ (the coefficients of the fixed effects), $\dot{u}_c$ (the coefficients of the random effects), $\dot{\Omega}$ (common covariance matrix of the random effects coefficients u_c) and $\dot{\sigma}_c^2$ (the variances of the residuals ϵ_c). Compared to Algorithm 3.1, the parameters $\dot{u}_c$ and $\dot{\Omega}$ are new.

We use the Gibbs sampler defined by Kasim and Raudenbush (1998, pp. 98–100) to create random draws from the parameters using a Markov chain Monte Carlo (MCMC) algorithm. The procedure repeats the following sampling steps:

$$\dot{\beta} \sim p(\beta | u_c, \sigma^2) \tag{3.12}$$
$$\dot{u}_c \sim p(u_c | \beta, \Omega, \sigma^2) \tag{3.13}$$
$$\dot{\Omega} \sim p(\Omega | u_c) \tag{3.14}$$
$$\dot{\sigma}_c^2 \sim p(\sigma^2 | \beta, u_c) \tag{3.15}$$

The exact specification of the prior and posterior distributions is fairly

Algorithm 3.7: Multilevel imputation with bootstrap under the linear normal model with class-wise error variances.♠

1. Draw a bootstrap sample of the level-2 units.
2. Within each level-2 unit, draw a bootstrap sample of the level-1 units.
3. Update $\dot{y}_{\text{obs}}$ and $\dot{X}_{\text{obs}}$ accordingly.
4. Estimate $\dot{\beta}$, $\dot{\sigma}_c^2$, and $\dot{\Omega}$ from $\dot{y}_{\text{obs}}$ and $\dot{X}_{\text{obs}}$ in the usual way, e.g., by `lme()` and `varIdent()` from the `nlme` package.
5. Sample $\dot{u}_c \sim p(u_c|\dot{\beta}, \dot{\Omega}, \dot{\sigma}^2)$ given the observed data y_{obs} and X_{obs}.
6. Calculate the n_0 total variances $\dot{s}^2 = (Z'_{\text{mis},[j]}\dot{\Omega}Z_{\text{mis},[j]} + \dot{\sigma}_c^2)$, where $j = 1, \ldots, n_0$ and c is the class to which j belongs.
7. Draw n_0 independent $N(0, 1)$ variates in vector $\dot{z}$.
8. Calculate the n_0 values $\dot{y} = X_{\text{mis}}\dot{\beta} + Z_{\text{mis}}\dot{u}_c + \dot{z}\dot{s}$, where c is the class to which j belongs.

complicated and will not be given here. The Bayesian method was implemented in `R` by Roel de Jong and is available in `mice` as the function `mice.impute.2L.norm()`.

The bootstrap method first draws a bootstrap sample for the complete data, fits the linear mixed effect model to it and calculates synthetic values according to the model estimates. This principle follows Algorithm 3.2.

Algorithm 3.7 first draws a sample from the level-2 units, followed by a second bootstrap sample of level-1 unit per drawn level-1 unit. One particular difficulty in this setup is that no random effects u_c are estimated for classes that are not drawn at step 1. We remedy this situation by drawing $\dot{u}_c$ from the original data, while fixing $\dot{\beta}, \dot{\Omega}, \dot{\sigma}^2$ at their estimates from the bootstrap sample. Once this is done, calculation of the synthetic values is straightforward.

3.8.4 Conclusion

Imputation of multilevel data is an area where work still remains to be done. In particular, we need faster algorithms and better coverage of and extensions to categorical data. On the other hand, the methodology that is currently available improves upon standard practice, and is therefore recommended over ad hoc alternatives.

3.9 Nonignorable missing data

3.9.1 Overview

All methods described thus far assume that the missing data mechanism is ignorable. In this case, there is no need for an explicit model of the missing data process (cf. Section 2.2.6). In reality, the mechanism may be nonignorable, even after accounting for any measurable factors that govern the response probability. In such cases, we can try to adapt the imputed data to make them more realistic. Since such adaptations are based on unverifiable assumptions, it is recommended to study carefully the impact of different possibilities on the final inferences by means of sensitivity analysis.

When is the assumption of ignorability suspect? It is hard to provide cut-and-dried criteria, but the following list illustrates some typical situations:

- If important variables that govern the missing data process are not available;
- If there is reason to believe that responders differ from non-responders, even after accounting for the observed information;
- If the data are truncated.

If ignorability does not hold, we need to model the distribution $P(Y, R)$ instead of $P(Y)$. For nonignorable missing data mechanisms, $P(Y, R)$ do not factorize into independent parts. Two main strategies to decompose $P(Y, R)$ are known as the *selection model* (Heckman, 1976) and the *pattern-mixture model* (Glynn et al., 1986). Little and Rubin (2002, ch. 15) and Little (2009) provide in-depth discussions of these models.

Imputations are created most easily under the pattern-mixture model. Herzog and Rubin (1983, pp. 222–224) proposed a simple and general family of nonignorable models that accounts for shift bias, scale bias and shape bias. Suppose that we expect that the nonrespondent data are shifted relative to the respondent data. Adding a simple shift parameter δ to the imputations creates a difference in the means of a δ. In a similar vein, if we suspect that the nonrespondents and respondents use different scales, we can multiply each imputation by a scale parameter. Likewise, if we suspect that the shapes of both distributions differ, we could redraw values from the candidate imputations with a probability proportional to the dissimilarity between the two distributions, a technique known as the SIR algorithm (Rubin, 1987b). We only discuss the shift parameter δ.

In practice, it may be difficult to specify the distribution of the nonrespondents, e.g., to provide a sensible specification of δ. One approach is to compare the results under different values of δ by sensitivity analysis. Though helpful, this puts the burden on the specification of realistic scenarios, i.e., a set of

Table 3.5: Numerical example of an nonignorable nonresponse mechanism, where more missing data occur in groups with lower blood pressures.

Y	$P(Y)$	$P(R=1\|Y)$	$P(Y\|R=1)$	$P(Y\|R=0)$
100	0.02	0.65	0.015	0.058
110	0.03	0.70	0.024	0.074
120	0.05	0.75	0.043	0.103
130	0.10	0.80	0.091	0.164
140	0.15	0.85	0.145	0.185
150	0.30	0.90	0.307	0.247
160	0.15	0.92	0.157	0.099
170	0.10	0.94	0.107	0.049
180	0.05	0.96	0.055	0.016
190	0.03	0.98	0.033	0.005
200	0.02	1.00	0.023	0.000
$\bar{Y}$	150.00		151.58	138.60

plausible δ-values. The next sections describe the selection model and pattern mixture in more detail, as a way to evaluate the plausibility of δ.

3.9.2 Selection model

The selection model (Heckman, 1976) decomposes the joint distribution $P(Y, R)$ as

$$P(Y, R) = P(Y)P(R|Y). \tag{3.16}$$

The selection model weights the marginal distribution $P(Y)$ in the population with the response weights $P(R|Y)$. Both $P(Y)$ and $P(R|Y)$ are unknown, and must be specified by the user. The model where $P(Y)$ is normal and where $P(R|Y)$ is a probit model is known as the Heckman model. This model is widely used in economics to correct for selection bias.

Numerical example. The column labeled Y in Table 3.5 contains the midpoints of 11 categories of systolic blood pressure. The column $P(Y)$ contains a hypothetically complete distribution of systolic blood pressure. It is specified here as symmetric with a mean of 150 mmHg (millimeters mercury). This distribution should be a realistic description of the *combined* observed and missing blood pressure values in the population of interest. The column $P(R = 1|Y)$ specifies the probability that blood pressure is actually observed at different levels of blood pressure. Thus, at a systolic blood pressure of 100 mmHg we expect that 65% of the data is observed. On the other hand, we expect that no missing data occur for those with a blood pressure of 200 mmHg. This specification produces 12.2% of missing data. The variability in the missingness probability is large, and reflects an extreme scenario where the missing data are created mostly at the lower blood pressures. Section 7.2.1

discusses why more missing data in the lower levels are plausible. When taken together, the columns $P(Y)$ and $P(R = 1|Y)$ specify a selection model.

3.9.3 Pattern-mixture model

The pattern-mixture model (Glynn et al., 1986; Little, 1993) decomposes the joint distribution $P(Y, R)$ as

$$\begin{aligned} P(Y, R) &= P(Y|R)P(R) && (3.17) \\ &= P(Y|R = 1)P(R = 1) + P(Y|R = 0)P(R = 0) && (3.18) \end{aligned}$$

Compared to Equation 3.16 this model only reverses the roles of Y and R, but the interpretation is quite different. The pattern-mixture model emphasizes that the combined distribution is a mix of the distributions of Y in the responders and nonresponders. The model needs a specification of the distribution $P(Y|R = 1)$ of the responders (which can be conveniently modeled after the data), and of the distribution $P(Y|R = 0)$ of the nonresponders (for which we have no data at all). The joint distribution is the mixture of these two distributions, with mixing probabilities $P(R = 1)$ and $P(R = 0) = 1 - P(R = 1)$, the overall proportions of observed and missing data, respectively.

Numerical example. The columns labeled $P(Y|R = 1)$ and $P(Y|R = 0)$ in Table 3.5 contain the probability per blood pressure category for the respondents and nonrespondents. Since more missing data are expected to occur at lower blood pressures, the mass of the nonresponder distribution has shifted toward the lower end of the scale. As a result, the mean of the nonresponder distribution is equal to 138.6 mmHg, while the mean of the responder distribution equals 151.58 mmHg.

3.9.4 Converting selection and pattern-mixture models

The pattern-mixture model and the selection model are connected via Bayes rule. Suppose that we have a mixture model specified as the probability distributions $P(Y|R = 0)$ and $P(Y|R = 1)$ plus the overall response probability $P(R)$. The corresponding selection model can be calculated as

$$P(R = 1|Y = y) = P(Y = y|R = 1)P(R = 1)/P(Y = y) \quad (3.19)$$

where the marginal distribution of Y is

$$P(Y = y) = P(Y = y|R = 1)P(R = 1) + P(Y = y|R = 0)P(R = 0) \quad (3.20)$$

Reversely, the pattern-mixture model can be calculated from the selection model as follows:

$$P(Y = y|R = r) = P(R = r|Y = y)P(Y = y)/P(R = r) \quad (3.21)$$

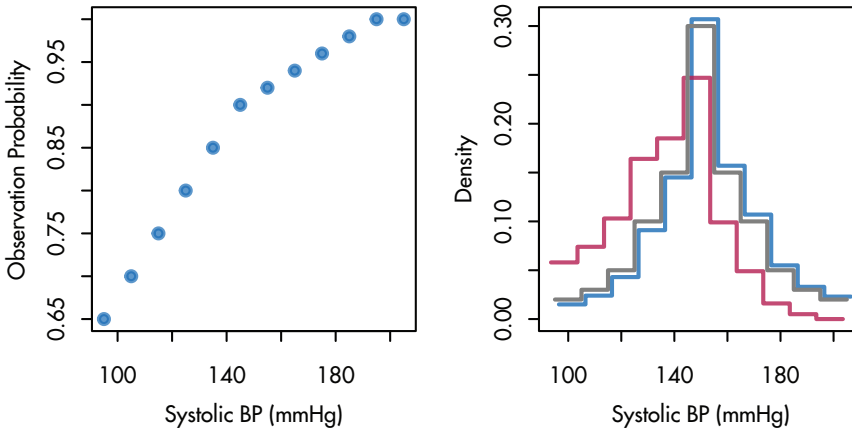


Figure 3.9: Graphic representation of the response mechanism for systolic bloood pressure in Table 3.5. See text for explanation.

where the overall probability of observed ($r = 1$) or missing ($r = 0$) data is equal to

$$P(R = r) = \sum_y P(R = r|Y = y)P(Y = y) \tag{3.22}$$

Numerical example. In Table 3.5 we calculate $P(Y = 100) = 0.015 \times 0.878 + 0.058 \times 0.122 = 0.02$. Likewise, we find $P(R = 1|Y) = 0.015 \times 0.878/0.02 = 0.65$ and $P(R = 0|Y) = 0.058 \times 0.122/0.02 = 0.35$. The reverse calculation is left as an exercise to the reader.

Figure 3.9 is a graphic illustration of the posited missing data mechanism. The left-hand figure displays the missingness probabilities $P(R|Y)$ of the selection model. The right-hand plot provides the distributions $P(Y|R)$ in the observed (blue) and missing (red) data in the corresponding pattern-mixture model. The hypothetically complete distribution is given by the black curve. The distribution of blood pressure in the group with missing blood pressures is quite different, both in form and location. At the same time, observe that the effect of missingness on the combined distribution is only slight. The reason is that 87% of the information is actually observed.

The mean of the distribution of the observed data remains almost unchanged (151.6 mmHg instead of 150 mmHg), but the mean of the distribution of the missing data is substantially lower at 138.6 mmHg. Thus, under the assumed selection model we expect that the mean of the imputed data should be $151.6 - 138.6 = 13$ mmHg lower than in the observed data.

Table 3.6: Difference between the means of the blood pressure distributions of the response and nonresponse groups, and its interpretation in the light of what we know about the data.

δ	Interpretation
0 mmHg	MCAR, δ too small
−5 mmHg	Small effect
−10 mmHg	Large effect
−15 mmHg	Extreme effect
−20 mmHg	Too extreme effect

3.9.5 Sensitivity analysis

Sections 3.9.2–3.9.4 provide different, though related, views on the assumed response model. A fairly extreme response model where the missingness probability increases from 0% to 35% in the outcome produces a mean difference of 13 mmHg. The effect in the combined distribution is much smaller: 1.6 mmHg.

Section 3.9.1 discussed the idea of adding some extra mmHg to the imputed values, a method known as δ-adjustment. It is important to form an idea of what reasonable values for δ could be. Under the posited model, $\delta = 0$ mmHg is clearly too small (as it assumes MCAR), whereas $\delta = -20$ mmHg is too extreme (as it can only occur if nearly all missing values occur in the lowest blood pressures). Table 3.6 provides an interpretation of various values for δ. The most likely scenarios would yield $\delta = -5$ or $\delta = -10$ mmHg.

In practice, part of δ may be realized through the predictors needed under MAR. It is useful to decompose δ as $\delta = \delta_{\text{MAR}} + \delta_{\text{MNAR}}$, where δ_{MAR} is the mean difference caused by the predictors in the imputation models, and where δ_{MNAR} is the mean difference caused by an additional nonignorable part of the imputation model. If candidate imputations are produced under MAR, we only need to add a constant δ_{MNAR}. Section 7.2 continues this application.

Adding a constant may seem overly simple, but it is actually quite powerful. In cases where no one model will be obviously more realistic than any other, Rubin (1987a, p. 203) stressed the need for easily communicated models, like a "20% increase over the ignorable value." Little (2009, p. 49) warned that it is easy to be enamored of complicated models for $P(Y, R)$ so that we may be "lulled into a false sense of complacency about the fundamental lack of identification," and suggested simple methods:

> The idea of adding offsets is simple, transparent, and can be readily accomplished with existing software.

Adding a constant or multiplying by a value are in fact the most direct ways to specify nonignorable models.

3.9.6 Role of sensitivity analysis

Nonignorable models are only useful after the possibilities to make the data "more MAR" have been exhausted. A first step is always to create the best possible imputation model based on the available data. Section 5.3.2 provides specific advice on how to build imputation models.

The MAR assumption has been proven defensible for intentional missing data. In general, however, we can never rule out the possibility that the data are MNAR. In order to cater for this possibility, many advise performing a sensitivity analysis on the final result. This is voiced most clearly in recommendation 15 of the National Research Council's advice on clinical trials (National Research Council, 2010):

> Recommendation 15: Sensitivity analysis should be part of the primary reporting of findings from clinical trials. Examining sensitivity to the assumptions about the missing data mechanism should be a mandatory component of reporting.

While there is much to commend this rule, we should refrain from doing sensitivity analysis just for the sake of it. The proper execution of a sensitivity analysis requires us to specify plausible scenarios. An extreme scenario like "suppose that all persons who leave the study die" can have a major impact on the study result, yet it could be highly improbable and therefore of limited interest.

Sensitivity analysis on factors that are already part of the imputation model is superfluous. Preferably, before embarking on a sensitivity analysis, there should be reasonable evidence that the MAR assumption is (still) inadequate after the available data have been taken into account. Such evidence is also crucial in formulating plausible MNAR mechanisms. Any decisions about scenarios for sensitivity analysis should be taken in discussion with subject-matter specialists. There is no purely statistical solution to the problem of nonignorable missing data. Sensitivity analysis can increase our insight into the stability of the results, but in my opinion we should only use it if we have a firm idea of which scenarios for the missingness would be reasonable.

In practice, we may lack such insights. In such instances, I would prefer a carefully constructed imputation model (which is based on all available data) over a poorly constructed sensitivity analysis.

3.10 Exercises

1. *MAR.* Reproduce Table 3.1 and Table 3.2 for MARRIGHT, MARMID and MARTAIL missing data mechanisms of Section 3.2.4.

   (a) Are there any choices that you need to make? If so, which?

(b) Consider the six possibilities to combine the missing data mechanism and missingness in x or y. Do you expect complete case analysis to perform well in each case?

(c) Do the Bayesian and bootstrap methods also work under the three MAR mechanisms?

2. *Parameter uncertainty.* Repeat the simulations of Section 3.2 on the `whiteside` data for different samples sizes.

(a) Use the method of Section 3.2.3 to generate an artificial population of 10000 synthetic gas consumption observations. Re-estimate the parameter from the artificial population. How close are they to the "true" values?

(b) Draw random samples from the artificial population. Systematically vary sample size. Is there some sample size at which "predict + noise" is as good as the Bayesian and bootstrap methods?

(c) Is the result identical for missing y and missing x?

(d) Is the result the same after including insulation status in the model?

3. *Tree imputation.* ♠ Write a function `mice.impute.tree()` that implements Algorithm 3.6 for a binary outcome. Use the function `rpart()` from the `rpart` package to fit the tree. Use the function `mice.impute.TF()` in Section 3.3.2 as a template. Provide an option `leavesize` to regulate the size of the terminal node.

(a) Were there any decisions you needed to make in programming `mice.impute.tree()`. If so, which?

(b) Do graphs demonstrate that the observed and imputed data match up as expected?

(c) Conduct a small simulation study using a complete data model that includes an interaction effect. Is your method unbiased?

(d) Is your method an improvement over predictive mean matching?

Chapter 4

Multivariate missing data

Chapter 3 dealt with univariate missing data. In practice, missing data may occur anywhere. This chapter discusses potential problems created by multivariate missing data, and outlines several approaches to deal with these issues.

4.1 Missing data pattern

4.1.1 Overview

Let the data be represented by the $n \times p$ matrix Y. In the presence of missing data Y is partially observed. Notation Y_j is the jth column in Y, and Y_{-j} indicates the complement of Y_j, that is, all columns in Y except Y_j. The *missing data pattern* of Y is the $n \times p$ binary response matrix R, as defined in Section 2.2.3.

For both theoretical and practical reasons, it is useful to distinguish various types of missing data patterns:

1. *Univariate and multivariate.* A missing data pattern is said to be univariate if there is only one variable with missing data.

2. *Monotone and non-monotone (or general).* A missing data pattern is said to be *monotone* if the variables Y_j can be ordered such that if Y_j is missing then all variables Y_k with $k > j$ are also missing. This occurs, for example, in longitudinal studies with drop-out. If the pattern is not monotone, it is called *non-monotone* or *general*.

3. *Connected and unconnected.* A missing data pattern is said to be *connected* if any observed data point can be reached from any other observed data point through a sequence of horizontal or vertical moves (like the rook in chess).

Figure 4.1 illustrates various data patterns in multivariate data. Monotone patterns can occur as a result of drop-out in longitudinal studies. If a pattern is monotone, the variables can be sorted conveniently according to the percentage of missing data. Univariate missing data form a special monotone pattern. Important computational savings are possible if the data are monotone.

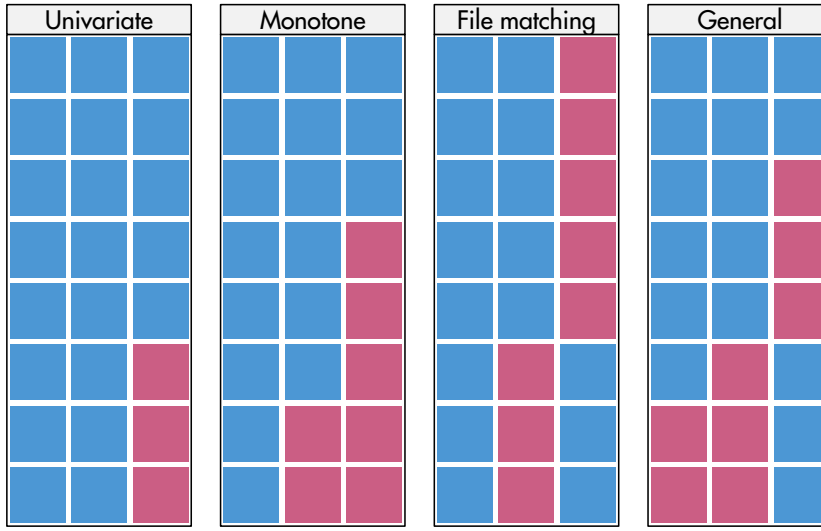


Figure 4.1: Some missing data patterns in multivariate data. Blue is observed, red is missing.

All patterns displayed in Figure 4.1 are connected. The file matching pattern is connected since it is possible to travel to all blue cells by horizontal or vertical moves. This pattern will become unconnected if we remove the first column. In contrast, after removing the first column from the general pattern in Figure 4.1 it is still connected through the first two rows.

Connected patterns are needed to identify unknown parameters. For example, in order to be able to estimate a correlation coefficient between two variables, they need to be connected, either directly by a set of cases that have scores on both, or indirectly through their relation with a third set of connected data. Unconnected patterns may arise in particular data collection designs, like data combination of different variables and samples, or potential outcomes.

Missing data patterns of longitudinal data organized in the "long format" (cf. Section 9.1) are more complex than the patterns in Figure 4.1. See Van Buuren (2011, p. 179) for some examples.

4.1.2 Summary statistics

The missing data pattern influences the amount of information that can be transferred between variables. Imputation can be more precise if other variables are non-missing for those cases that are to be imputed. The reverse is

also true. Predictors are potentially more powerful if they have are non-missing in rows that are vastly incomplete. This section discusses various measures of the missing data pattern.

The function `md.pattern()` in `mice` calculates the frequencies of the missing data patterns. For example, the frequency pattern of the dataset `pattern4` in Figure 4.1 is

```
> md.pattern(pattern4)
  A B C
2 1 1 1 0
1 1 0 1 1
3 1 1 0 1
2 0 0 1 2
  2 3 3 8
```

The columns `A`, `B` and `C` are either 0 (missing) or 1 (observed). The first column provides the frequency of each pattern. The last column lists the number of missing entries per pattern. The bottom row provides the number of missing entries per variable, and the total number of missing cells. In practice, `md.pattern()` is primarily useful for datasets with a small number of columns.

Alternative measures start from pairs of variables. A pair of variables (Y_j, Y_k) can have four missingness patterns:

1. both Y_j and Y_k are observed (pattern `rr`);
2. Y_j is observed and Y_k is missing (pattern `rm`);
3. Y_j is missing and Y_k is observed (pattern `mr`);
4. both Y_j and Y_k are missing (pattern `mm`).

For example, for the monotone pattern in Figure 4.1 the frequencies are:

```
> p <- md.pairs(pattern4)
> p
$rr
  A B C
A 6 5 3
B 5 5 2
C 3 2 5

$rm
  A B C
A 0 1 3
B 0 0 3
C 2 3 0
```

```
$mr
  A B C
A 0 0 2
B 1 0 3
C 3 3 0

$mm
  A B C
A 2 2 0
B 2 3 0
C 0 0 3
```

Thus, for pair `(A,B)` there are five completely observed pairs (in `rr`), no pairs in which `A` is observed and `B` missing (in `rm`), one pair in which `A` is missing and `B` is observed (in `mr`) and two pairs with both missing `A` and `B`. Note that these numbers add up to the total sample size.

The *proportion of usable cases* (Van Buuren et al., 1999) for imputing variable Y_j from variable Y_k is defined as

$$I_{jk} = \frac{\sum_i^n (1 - r_{ij}) r_{ik}}{\sum_i^n 1 - r_{ij}} \tag{4.1}$$

This quantity can be interpreted as the number of pairs (Y_j, Y_k) with Y_j missing and Y_k observed, divided by the total number of missing cases in Y_j. The proportion of usable cases I_{jk} equals 1 if variable Y_k is observed in all records where Y_j is missing. The statistic can be used to quickly select potential predictors Y_k for imputing Y_j based on the missing data pattern. High values of I_{jk} are preferred. For example, we can calculate I_{jk} in the dataset `pattern4` in Figure 4.1 for all pairs (Y_j, Y_k) by

```
> p$mr/(p$mr + p$mm)
      A B C
A 0.000 0 1
B 0.333 0 1
C 1.000 1 0
```

The first row contains $I_{\text{AA}} = 0$, $I_{\text{AB}} = 0$ and $I_{\text{AC}} = 1$. This informs us that `B` is not relevant for imputing `A` since there are no observed cases in `B` where `A` is missing. However, `C` is observed for both missing entries in `A`, and may thus be a relevant predictor. The I_{jk} statistic is an *inbound* statistic that measures how well the missing entries in variable Y_j are connected to the rest of the data.

The *outbound* statistic O_{jk} measures how well observed data in variable Y_j connect to missing data in the rest of the data. The statistic is defined as

$$O_{jk} = \frac{\sum_i^n r_{ij} (1 - r_{ik})}{\sum_i^n r_{ij}} \tag{4.2}$$

This quantity is the number of observed pairs (Y_j, Y_k) with Y_j observed and Y_k missing, divided by the total number of observed cases in Y_j. The quantity O_{jk} equals 1 if variable Y_j is observed in all records where Y_k is missing. The statistic can be used to evaluate whether Y_j is a potential predictor for imputing Y_k. We can calculate O_{jk} in the dataset `pattern4` in Figure 4.1 for all pairs (Y_j, Y_k) by

```
> p$rm/(p$rm + p$rr)
    A     B   C
A 0.0 0.167 0.5
B 0.0 0.000 0.6
C 0.4 0.600 0.0
```

Thus `A` is potentially more useful to impute `C` (3 out of 6) than `B` (1 out of 6).

4.1.3 Influx and outflux

The inbound and outbound statistics in the previous section are defined for variable pairs (Y_j, Y_k). This section describes two overall measures of how each variable connects to others: influx and outflux.

The *influx coefficient* I_j is defined as

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n r_{ik}} \tag{4.3}$$

The coefficient is equal to the number of variable pairs (Y_j, Y_k) with Y_j missing and Y_k observed, divided by the total number of observed data cells. The value of I_j depends on the proportion of missing data of the variable. Influx of a completely observed variable is equal to 0, whereas for completely missing variables we have $I_j = 1$. For two variables with the same proportion of missing data, the variable with higher influx is better connected to the observed data, and might thus be easier to impute.

The *outflux coefficient* O_j is defined in an analogous way as

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n r_{ij} (1 - r_{ik})}{\sum_k^p \sum_i^n 1 - r_{ij}} \tag{4.4}$$

The quantity O_j is the number of variable pairs with Y_j observed and Y_k missing, divided by the total number of incomplete data cells. Outflux is an indicator of the potential usefulness of Y_j for imputing other variables. Outflux depends on the proportion of missing data of the variable. Outflux of a completely observed variable is equal to 1, whereas outflux of a completely missing variable is equal to 0. For two variables having the same proportion of missing data, the variable with higher outflux is better connected to the missing data, and thus potentially more useful for imputing other variables.

The function `flux()` in `mice` calculates I_j and O_j for all variables. For example, for `pattern4` we obtain

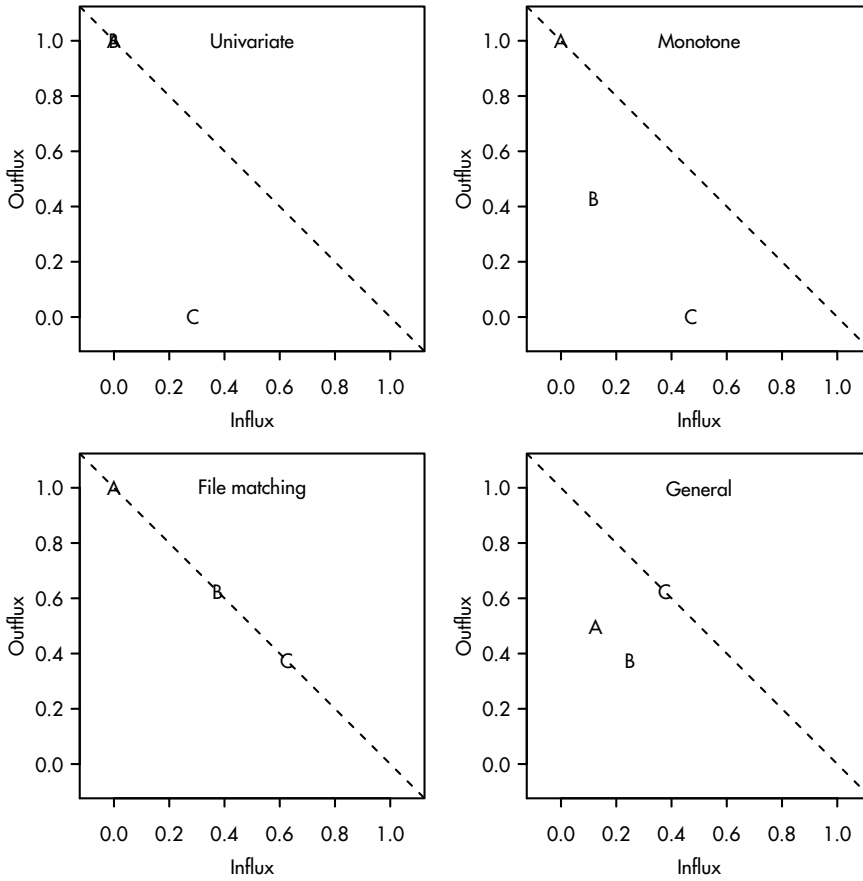


Figure 4.2: Fluxplot: Outflux versus influx in the four missing data patterns from Figure 4.1. The influx of a variable quantifies how well its missing data connect to the observed data on other variables. The outflux of a variable quantifies how well its observed data connect to the missing data on other variables. In general, higher influx and outflux values are preferred.

```
> flux(pattern4)[, 1:3]
   pobs influx outflux
A 0.750  0.125   0.500
B 0.625  0.250   0.375
C 0.625  0.375   0.625
```

The rows correspond to the variables. The columns contain the proportion of observed data, I_j and O_j. Figure 4.2 shows the influx-outflux pattern of the four patterns in Figure 4.1 produced by `fluxplot()`. In general, variables that are located higher up in the display are more complete and thus

potentially more useful for imputation. It is often (but not always) true that $I_j + O_j \leq 1$, so in practice variables closer to the subdiagonal are typically better connected than those farther away. The fluxplot can be used to spot variables that clutter the imputation model. Variables that are located in the lower regions (especially near the lower-left corner) *and* that are uninteresting for later analysis are better removed from the data prior to imputation.

Influx and outflux are summaries of the missing data pattern intended to aid in the construction of imputation models. Keeping everything else constant, variables with high influx and outflux are preferred. Realize that outflux indicates the potential (and not actual) contribution to impute other variables. A variable with high O_j may turn out to be useless for imputation if it is unrelated to the incomplete variables. On the other hand, the usefulness of a highly predictive variable is severely limited by a low O_j. More refined measures of usefulness are conceivable, e.g., multiplying O_j by the average proportion of explained variance. Also, we could specialize to one or a few key variables to impute. Alternatively, analogous measures for I_j could be useful. The further development of diagnostic summaries for the missing data pattern is a promising area for further investigation.

4.2 Issues in multivariate imputation

Most imputation models for Y_j use the remaining columns Y_{-j} as predictors. The rationale is that conditioning on Y_{-j} preserves the relations among the Y_j in the imputed data. Van Buuren et al. (2006) identified various practical problems that can occur in multivariate missing data:

- The predictors Y_{-j} themselves can contain missing values;
- "Circular" dependence can occur, where Y_j^{mis} depends on Y_h^{mis} and Y_h^{mis} depends on Y_j^{mis} with $h \neq j$, because in general Y_j and Y_h are correlated, even given other variables;
- Variables are often of different types (e.g., binary, unordered, ordered, continuous), thereby making the application of theoretically convenient models, such as the multivariate normal, theoretically inappropriate;
- Especially with large p and small n, collinearity or empty cells can occur;
- The ordering of the rows and columns can be meaningful, e.g., as in longitudinal data;
- The relation between Y_j and predictors Y_{-j} can be complex, e.g., nonlinear, or subject to censoring processes;

- Imputation can create impossible combinations, such as pregnant fathers.

This list is by no means exhaustive, and other complexities may appear for particular data. The next sections discuss three general strategies for imputing multivariate data:

- *Monotone data imputation*. For monotone missing data patterns, imputations are created by a sequence of univariate methods;
- *Joint modeling*. For general patterns, imputations are drawn from a multivariate model fitted to the data;
- *Fully conditional specification*, also known as *chained equations* and *sequential regressions*. For general patterns, a multivariate model is implicitly specified by a set of conditional univariate models. Imputations are created by drawing from iterated conditional models.

4.3 Monotone data imputation

4.3.1 Overview

Imputations of monotone missing data can be generated by specifying a sequence of univariate methods (one for each incomplete column), followed by drawing sequentially synthetic observations under each method.

Suppose that variables $Y_1, \ldots, Y_p$ are ordered into a monotone missing data pattern. The general recommended procedure is as follows (Rubin, 1987a, p. 172). The missing values of Y_1 are imputed from a (possibly empty) set of complete covariates X ignoring $Y_2, \ldots, Y_p$. Next, the missing values of Y_2 are imputed from (Y_1, X) ignoring $Y_3, \ldots, Y_p$, and so on. The procedure ends after Y_p is imputed from $(X, Y_1, \ldots, Y_{p-1})$. The univariate imputation methods as discussed in Chapter 3 can be used as building blocks. For example, Y_1 can be imputed by logistic regression, Y_2 by predictive mean matching, and so on.

Numerical example. The first three columns of the data frame `nhanes2` in `mice` have a monotone missing data pattern. In terms of the above notation, X contains the complete variable `age`, Y_1 is the variable `hyp`, and Y_2 is the variable `bmi`. Monotone data imputation can be applied to generate $m = 2$ complete datasets by:

```
> data <- nhanes2[, 1:3]
> md.pattern(data)
   age hyp bmi
16   1   1   1  0
 1   1   1   0  1
```

```
 8   1   0   0  2
     0   8   9 17
> imp <- mice(data, visit = "monotone", maxit = 1,
      m = 2)
 iter imp variable
  1   1  hyp  bmi
  1   2  hyp  bmi
```

The `md.pattern()` function outputs the three available data patterns in `data`. There are 16 complete rows, one row with missing `bmi`, and eight rows where both `bmi` and `hyp` are missing. The argument `visit="monotone"` specifies that the visit sequence should be equal to the number of missing data per variable (so first `hyp` and then `bmi`). Since one iteration is enough, we use `maxit=1` to limit the calculations. This code imputes `hyp` by logistic regression and `bmi` by predictive mean matching, the default methods for binary and continuous data, respectively.

Monotone data imputation requires that the missing data pattern is monotone. In addition, there is a second, more technical requirement: the parameters of the imputation models should be *distinct* (Rubin, 1987a, pp. 174–178). Let the jth imputation model be denoted by $P(Y_j^{\text{mis}}|X, Y_1, \ldots, Y_{p-1}, \phi_j)$, where ϕ_j represents the unknown parameters of the imputation model. For valid likelihood inferences, $\phi_1, \ldots, \phi_p$ should be distinct in the sense that the parameter space $\phi = (\phi_1, \ldots, \phi_p)$ in the multivariate model for the data is the cross-product of the individual parameter spaces (Schafer, 1997, p. 219). For Bayesian inference, it is required that the prior density of all parameters $\pi(\phi)$ factors into p independent densities $\pi(\phi) = \pi_1(\phi_1)\pi_2(\phi_2), \ldots, \pi_p(\phi_p)$ (Schafer, 1997, p. 224). In most applications these requirements are unlikely to limit the practical usefulness of the method because the parameters are typically unrelated and allowed to vary freely. We need to be aware, however, that monotone data imputation may fail if the parameters of imputation models for different Y_j somehow depend on each other.

4.3.2 Algorithm

Algorithm 4.1 provides the main steps of monotone data imputation. We order the variables according to their missingness, and impute from left to right. In practice, a pair of "draw-impute" steps is executed by one of the univariate methods of Chapter 3. Both Bayesian and bootstrap imputation methods can be used, and can in fact be mixed. There is no need to iterate, and convergence is immediate. The algorithm is replicated m times from different starting points to obtain m multiply imputed datasets.

Monotone data imputation is fast and flexible, but requires a monotone pattern. In practice, a dataset may be near-monotone, and may become monotone if a small fraction of the missing data were imputed. For example, some subjects may drop out of the study resulting in a monotone pattern. There

Algorithm 4.1: Monotone data imputation of multivariate missing data.♠

1. Sort the data Y_j^{obs} with $j = 1, \ldots, p$ according to their missingness.
2. Draw $\dot{\phi}_1 \sim P(Y_1^{\text{obs}}|X)$.
3. Impute $\dot{Y}_1 \sim P(Y_1^{\text{mis}}|X, \dot{\phi}_1)$.
4. Draw $\dot{\phi}_2 \sim P(Y_2^{\text{obs}}|X, \dot{Y}_1)$.
5. Impute $\dot{Y}_2 \sim P(Y_2^{\text{mis}}|X, \dot{Y}_1, \dot{\phi}_2)$.
6. $\vdots$
7. Draw $\dot{\phi}_p \sim P(Y_p^{\text{obs}}|X, \dot{Y}_1, \ldots, \dot{Y}_{p-1})$.
8. Impute $\dot{Y}_p \sim P(Y_p^{\text{mis}}|X, \dot{Y}_1, \ldots, \dot{Y}_{p-1}, \dot{\phi}_p)$.

could be some unplanned missing data that destroy the monotone pattern. In such cases it can be computationally efficient to impute the data in two steps. First, fill in the missing data in a small portion of the data to restore the monotone pattern, and then apply the monotone data imputation (Li, 1988; Rubin and Schafer, 1990; Liu, 1993; Schafer, 1997; Rubin, 2003). There are often more ways to impute toward monotonicity, so a choice is necessary. Rubin and Schafer (1990) suggested ordering the variables according to the missing data rate.

Numerical example. The `nhanes2` data in `mice` contains 3 out of 27 missing values that destroy the monotone pattern: one for `hyp` (in row 6) and two for `bmi` (in rows 3 and 6). An approximate version of the above two-step algorithm can be executed as

```
> ini <- mice(nhanes2, maxit = 0)
> pred <- ini$pred
> pred["bmi", "chl"] <- 0
> pred["hyp", c("chl", "bmi")] <- 0
> imp <- mice(nhanes2, vis = "monotone", pred = pred,
      maxit = 1, m = 2)
 iter imp variable
  1   1  hyp  bmi  chl
  1   2  hyp  bmi  chl
```

This code actually combines both steps by sequentially imputing `hyp`, `bmi` and `chl` in one pass through the data. The first four statements change the default predictor matrix of `mice()` to reflect the visit sequence. The code yields valid imputations for the 24 missing values that belong to the monotone part.

Observe that the imputed values for the missing `hyp` value in row 3 could also depend on `bmi` and `chl`, but under monotone data imputation both variables do not appear in the model for `hyp`. In principle, we can improve the method by incorporating `bmi` and `chl` into the model and iterating. We explore this technique in more detail in Section 4.5, but first we study an alternative.

4.4 Joint modeling

4.4.1 Overview

Joint modeling (JM) starts from the assumption that the data can be described by a multivariate distribution. Assuming ignorability, imputations are created as draws from the fitted distribution. The model can be based on any multivariate distribution. The multivariate normal distribution is most widely applied.

The general idea is as follows. For a general missing data pattern, missing data can occur anywhere in Y, so in practice the distribution from which imputations are to be drawn varies from row to row. For example, if the missingness pattern of row i is $r_{[i]} = (0, 0, 1, 1)$, then we need to draw imputations from the bivariate distribution $P_i(Y_1^{\text{mis}}, Y_2^{\text{mis}}|Y_3, Y_4, \phi_{1,2})$, whereas if $r_{[i']} = (0, 1, 1, 1)$ we need draws from the univariate distribution $P_{i'}(Y_1^{\text{mis}}|Y_1, Y_3, Y_4, \phi_1)$.

4.4.2 Continuous data ♠

Under the assumption of multivariate normality $Y \sim N(\mu, \Sigma)$, the ϕ-parameters of these imputation models are functions of $\theta = (\mu, \Sigma)$ (Schafer, 1997, p. 157). The *sweep operator* transforms θ into ϕ by converting outcome variables into predictors, while the *reverse sweep operator* allows for the inverse operation (Beaton, 1964). The sweep operators allow rapid calculation of the ϕ parameters for imputation models that pertain to different missing data patterns. For reasons of efficiency, rows can be grouped along the missing data pattern. See Little and Rubin (2002, pp. 148–156) and Schafer (1997, p. 157–163) for computational details.

The θ-parameters are usually unknown. For non-monotone missing data, however, it is generally difficult to estimate θ from Y_{obs} directly. The solution is to iterate imputation and parameter estimation using a general algorithm known as *data augmentation* (Tanner and Wong, 1987). At step t, the algorithm draws Y_{mis} and θ by alternating the following steps:

$$\dot{Y}_{\text{mis}}^t \sim P(Y_{\text{mis}}|Y_{\text{obs}}, \dot{\theta}^{t-1}) \tag{4.5}$$

$$\dot{\theta}^t \sim P(\theta|Y_{\text{obs}}, \dot{Y}_{\text{mis}}^t) \tag{4.6}$$

Algorithm 4.2: Imputation of missing data by a joint model for multivariate normal data.♠

1. Sort the rows of Y into S missing data patterns $Y_{[s]}, s = 1, \ldots, S$.
2. Initialize $\theta^0 = (\mu^0, \Sigma^0)$ by a reasonable starting value.
3. Repeat for $t = 1, \ldots, T$:
4. Repeat for $s = 1, \ldots, S$:
5. Calculate parameters $\dot{\phi}_s = \text{SWP}(\dot{\theta}^{t-1}, s)$ by sweeping the predictors of pattern s out of $\dot{\theta}^{t-1}$.
6. Calculate p_s as the number missing data in pattern s. Calculate $o_s = p - p_s$.
7. Calculate the Choleski decomposition C_s of the $p_s \times p_s$ submatrix of $\dot{\phi}_s$ corresponding to the missing data in pattern s.
8. Draw a random vector $z \sim N(0, 1)$ of length p_s.
9. Take $\dot{\beta}_s$ as the $o_s \times p_s$ submatrix of $\dot{\phi}_s$ of regression weights.
10. Calculate imputations $\dot{Y}^t_{[s]} = Y^{\text{obs}}_{[s]}\dot{\beta}_s + C'_s z$, where $Y^{\text{obs}}_{[s]}$ is the observed data in pattern s.
11. End repeat s.
12. Draw $\dot{\theta}^t = (\dot{\mu}, \dot{\Sigma})$ from the normal inverted-Wishart distribution according to Schafer (1997, p. 184).
13. End repeat t.

where imputations from $P(Y_{\text{mis}}|Y_{\text{obs}}, \dot{\theta}^{t-1})$ are drawn by the method as described in the previous section, and where draws from the parameter distribution $P(\theta|Y_{\text{obs}}, \dot{Y}^t_{\text{mis}})$ are generated according to the method of Schafer (1997, p. 184).

Algorithm 4.2 lists the major steps needed to impute multivariate missing data under the normal model. Additional background can be found in Li (1988), Rubin and Schafer (1990) and Schafer (1997). Song and Belin (2004) generated multiple imputations under the common factor model. The performance of the method was found to be similar to that of the multivariate normal distribution, the main pitfall being the danger of setting the numbers of factors too low.

Schafer (1997, p. 211–218) reported simulations that showed that imputations generated under the multivariate normal model are robust to non-normal

data. Demirtas et al. (2008) confirmed this claim in a more extensive simulation study. The authors conclude that "imputation under the assumption of normality is a fairly reasonable tool, even when the assumption of normality is clearly violated; the fraction of missing information is high, especially when the sample size is relatively large." It is often beneficial to transform the data before imputation toward normality, especially if the scientifically interesting parameters are difficult to estimate, like quantiles or variances. For example, we could apply a logarithmic transformation before imputation to remove skewness, and apply an exponential transformation after imputation to revert to the original scale. Some work on automatic transformation methods for joint models is available. Van Buuren et al. (1993) developed an iterative transformation-imputation algorithm that finds optimal transformations of the variables toward multivariate normality. The algorithm is iterative because the multiply imputed values contribute to define the transformation, and vice versa. Transformations toward normality have also been incorporated in `transcan()` and `aregImpute()` of the `Hmisc` package in `R` (Harrell, 2001).

4.4.3 Categorical data

The multivariate normal model is often applied to categorical data. Schafer (1997, p. 148) suggested rounding off continuous imputed values in categorical data to the nearest category "to preserve the distributional properties as fully as possible and to make them intelligible to the analyst." This advice was questioned by Horton et al. (2003), who showed that simple rounding may introduce bias in the estimates of interest, in particular for binary variables. Allison (2005) found that it is usually better not to round the data, and preferred methods specifically designed for categorical data, like logistic regression imputation or discriminant analysis imputation. Bernaards et al. (2007) confirmed the results of Horton et al. (2003) for simple rounding, and proposed two improvements to simple rounding: *coin flip* and *adaptive rounding*. Their simulations showed that "adaptive rounding seemed to provide the best performance, although its advantage over simple rounding was sometimes slight." Further work has been done by Yucel et al. (2008), who proposed rounding such that the marginal distribution in the imputations is similar to that of the observed data. Alternatively, Demirtas (2009) proposed two rounding methods based on logistic regression and an additional drawing step that makes rounding dependent on other variables in the imputation model. A single best rounding rule for categorical data has yet to be identified. Demirtas (2010) encourages researchers to avoid rounding altogether, and apply methods specifically designed for categorical data.

Several joint models for categorical variables have been proposed that do not rely on rounding. Schafer (1997) proposed several techniques to impute categorical data and mixed continuous-categorical data. Missing data in contingency tables can be imputed under the log-linear model. The model preserves higher-order interactions, and works best if the number of variables is

small, say, up to six. Mixed continuous-categorical data can be imputed under the general location model originally developed by Olkin and Tate (1961). This model combines the log-linear and multivariate normal models by fitting a restricted normal model to each cell of the contingency table. Further extensions have been suggested by Liu and Rubin (1998) and Peng et al. (2004). Belin et al. (1999) pointed out some limitations of the general location model for a larger dataset with 16 binary and 18 continuous variables. Their study found substantial differences between the imputed and follow-up data, especially for the binary data.

Alternative imputation methods based on joint models have been developed. Van Buuren and Van Rijckevorsel (1992) maximized internal consistency by the k-means clustering algorithm, and outlined methods to generate multiple imputations. Van Ginkel et al. (2007) proposed *two-way imputation*, a technique for imputing incomplete categorical data by conditioning on the row and column sum scores of the multivariate data. This method has applications for imputing missing test item responses. Vermunt et al. (2008) developed an imputation method based on latent class analysis. Latent class models can be used to describe complex association structures, like in log-linear analysis, but in addition they allow for larger and more flexible imputation models. Goldstein et al. (2009) described a joint model for mixed continuous-categorical data with a multilevel structure. Chen et al. (2011) proposed a class of models that specifies the conditional density by an odds ratio representation relative to the center of the distribution. This allows for separate models of the odds ratio function and the conditional density at the center.

4.5 Fully conditional specification

4.5.1 Overview

Fully conditional specification (FCS) (Van Buuren et al., 2006; Van Buuren, 2007a) imputes multivariate missing data on a variable-by-variable basis. The method requires a specification of an imputation model for each incomplete variable, and creates imputations per variable in an iterative fashion.

In contrast to joint modeling, FCS specifies the multivariate distribution $P(Y, X, R|\theta)$ through a set of conditional densities $P(Y_j|X, Y_{-j}, R, \phi_j)$. This conditional density is used to impute Y_j given X, Y_{-j} and R. Starting from simple random draws from the marginal distribution, imputation under FCS is done by iterating over the conditionally specified imputation models. The methods of Chapter 3 may act as building blocks. FCS is a natural generalization of univariate imputation.

Rubin (1987a, pp. 160–166) subdivided the work needed to create imputations into three tasks. The *modeling task* chooses a specific model for the data,

the *estimation task* formulates the posterior parameters distribution given the model and the *imputation task* takes a random draws for the missing data by drawing successively from parameter and data distributions. FCS directly specifies the conditional distributions from which draws should be made, and hence bypasses the need to specify a multivariate model for the data.

The idea of conditionally specified models is quite old. Conditional probability distributions follow naturally from the theory of stochastic Markov chains (Bartlett, 1978, pp. 34–41, pp. 231–236). In the context of spatial data, Besag preferred the use of conditional probability models over joint probability models, since "the conditional probability approach has greater intuitive appeal to the practising statistician" (Besag, 1974, p. 223).

In the context of missing data imputation, similar ideas have surfaced under a variety of names: stochastic relaxation (Kennickell, 1991), variable-by-variable imputation (Brand, 1999), switching regressions (Van Buuren et al., 1999), sequential regressions (Raghunathan et al., 2001), ordered pseudo-Gibbs sampler (Heckerman et al., 2001), partially incompatible MCMC (Rubin, 2003), iterated univariate imputation (Gelman, 2004), chained equations (Van Buuren and Groothuis-Oudshoorn, 2000) and fully conditional specification (FCS) (Van Buuren et al., 2006).

4.5.2 The MICE algorithm

There are several ways to implement imputation under conditionally specified models. Algorithm 4.3 describes one particular instance: the MICE algorithm (Van Buuren and Groothuis-Oudshoorn, 2000, 2011). The algorithm starts with a random draw from the observed data, and imputes the incomplete data in a variable-by-variable fashion. One iteration consists of one cycle through all Y_j. The number of iterations T can often be low, say 5 or 10. The MICE algorithm generates multiple imputations by executing Algorithm 4.3 in parallel m times.

The MICE algorithm is a Markov chain Monte Carlo (MCMC) method, where the state space is the collection of all imputed values. More specifically, if the conditionals are compatible (cf. Section 4.5.4), the MICE algorithm is a Gibbs sampler, a Bayesian simulation technique that samples from the conditional distributions in order to obtain samples from the joint distribution (Gelfand and Smith, 1990; Casella and George, 1992). In conventional applications of the Gibbs sampler the full conditional distributions are derived from the joint probability distribution (Gilks, 1996). In the MICE algorithm, the conditional distributions are under direct control of the user, and so the joint distribution is only implicitly known, and may not actually exist. While the latter is clearly undesirable from a theoretical point of view (since we do not know the joint distribution to which the algorithm converges), in practice it does not seem to hinder useful applications of the method (cf. Section 4.5.4).

In order to converge to a stationary distribution, a Markov chain needs to satisfy three important properties (Roberts, 1996; Tierney, 1996):

Algorithm 4.3: MICE algorithm for imputation of multivariate missing data.♠

1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, T$:
4. Repeat for $j = 1, \ldots, p$:
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

- *irreducible*, the chain must be able to reach all interesting parts of the state space;
- *aperiodic*, the chain should not oscillate between different states;
- *recurrence*, all interesting parts can be reached infinitely often, at least from almost all starting points.

Do these properties hold for the MICE algorithm? Irreducibility is generally not a problem since the user has large control over the interesting parts of the state space. This flexibility is actually the main rationale for FCS instead of a joint model.

Periodicity is a potential problem, and can arise in the situation where imputation models are clearly inconsistent. A rather artificial example of an oscillatory behavior occurs when Y_1 is imputed by $Y_2\beta + \epsilon_1$ and Y_2 is imputed by $-Y_1\beta + \epsilon_2$ for some fixed, nonzero β. The sampler will oscillate between two qualitatively different states, so the correlation between Y_1 and Y_2 after imputing Y_1 will differ from that after imputing Y_2. In general, we would like the statistical inferences to be independent of the stopping point. A way to diagnose the *ping-pong* problem is to stop the chain at different points. The stopping point should not affect the statistical inferences. The addition of noise to create imputations is a safeguard against periodicity, and allows the sampler to "break out" more easily.

Non-recurrence may also be a potential difficulty, manifesting itself as explosive or non-stationary behavior. For example, if imputations are made through deterministic functions, the Markov chain may lock up. Such cases can sometimes be diagnosed from the trace lines of the sampler. See Section 5.5.2 for an example. My experience is that as long as the parameters of imputation models are estimated from the data, non-recurrence is mild or absent.

The required properties of the MCMC method can be translated into conditions on the eigenvalues of the matrix of transition probabilities (MacKay, 2003, pp. 372–373). The development of practical tools that put these conditions to work for multiple imputation is still an ongoing research problem.

4.5.3 Performance

Each conditional density has to be specified separately, so FCS requires some modeling effort on the part of the user. Most software provides reasonable defaults for standard situations, so the actual effort required may be small.

A number of simulation studies provide evidence that FCS generally yields estimates that are unbiased and that possess appropriate coverage (Brand, 1999; Raghunathan et al., 2001; Brand et al., 2003; Tang et al., 2005; Van Buuren et al., 2006; Horton and Kleinman, 2007; Yu et al., 2007).

4.5.4 Compatibility ♠

Gibbs sampling is based on the idea that knowledge of the conditional distributions is sufficient to determine a joint distribution, if it exists. Two conditional densities $p(Y_1|Y_2)$ and $p(Y_2|Y_1)$ are said to be *compatible* if a joint distribution $p(Y_1, Y_2)$ exists that has $p(Y_1|Y_2)$ and $p(Y_2|Y_1)$ as its conditional densities. More precisely, the two conditional densities are compatible if and only if their density ratio $p(Y_1|Y_2)/p(Y_2|Y_1)$ factorizes into the product $u(Y_1)v(Y_2)$ for some integrable functions u and v (Besag, 1974). So, the joint distribution either exists and is unique, or does not exist.

What happens when the joint distribution does not exist? The MICE algorithm is ignorant of the non-existence of the joint distribution, and happily produces imputations whether the joint distribution exists or not. However, can the imputed data be trusted when we cannot find a joint distribution $p(Y_1, Y_2)$ that has $p(Y_1|Y_2)$ and $p(Y_2|Y_1)$ as its conditionals?

In practice, incompatibility issues may arise in MICE if deterministic functions of the data are imputed along with their originals. For example, the imputation model may contain interaction terms, data summaries or nonlinear functions of the data. Such terms may introduce feedback loops and impossible combinations into the system, which can invalidate the imputations (Van Buuren and Groothuis-Oudshoorn, 2011). It is important to diagnose this behavior, and eliminate feedback loops from the system. Chapter 5 describes the tools to do this.

Van Buuren et al. (2006) contains a small simulation study using strongly

incompatible models. The adverse effects on the estimates after multiple imputation were only minimal in the cases studied. Though FCS is only guaranteed to work if the conditionals are compatible, in practice it appears that it is robust when this condition is not met. More work is needed to verify this claim in more general and more realistic settings.

If the joint density itself is of genuine scientific interest, we should carefully evaluate the effect that imputations might have on the estimate of the distribution. For example, incompatible conditionals could produce a ridge (or spike) in an otherwise smooth density, and the location of the ridge may actually depend on the stopping point. If such is the case, then we should have a reason to favor a particular stopping point. Alternatively, we might try to reformulate the imputation model so that the stopping point effect disappears.

In the majority of cases, however, scientific interest will focus on quantities that are more remote to the joint density, such as regression weights, factor loadings, prevalence estimates, and so on. In such cases, the joint distribution is more like a nuisance factor that has no intrinsic value. MICE attempts to produce synthetic values that look sensible and that preserve the relations in the data. Gelman (2004) argues that

> having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g., zero/nonzero features in income components, bounds, skip patterns, nonlinear relations, interactions).

Apart from potential feedback problems, it appears that incompatibility is a relatively minor problem in practice, especially if the missing data rate is modest.

Arnold and Press (1989) provide necessary and sufficient conditions for the existence of a joint distribution given two conditional densities. See Arnold et al. (1999) for multivariate extensions. Gelman and Speed (1993) concentrate on the question whether an arbitrary mix of conditional and marginal distribution yields a unique joint distribution. Further theoretical work has been done by Arnold et al. (2002). The field has recently become very active. Several methods for identifying compatibility from actual data have been developed in the last few years (Tian et al., 2009; Ip and Wang, 2009; Tan et al., 2010; Wang and Kuo, 2010; Kuo and Wang, 2011; Chen, 2011). It is not yet known how well these methods will work in the context of missing data.

4.5.5 Number of iterations

When m sampling streams are calculated in parallel, monitoring convergence is done by plotting one or more statistics of interest in each stream against iteration number t. Common statistics to be plotted are the mean and standard deviation of the synthetic data, as well as the correlation between different variables. The pattern should be free of trend, and the variance within a chain should approximate the variance between chains.

In practice, a low number of iterations appears to be enough. Brand (1999) and (Van Buuren et al., 1999) set the number of iterations T quite low, usually somewhere between 5 to 20 iterations. This number is much lower than in other applications of MCMC methods, which often require thousands of iterations.

Why can the number of iterations in MICE be so low? First of all, realize that the imputed data $\dot{Y}_{\text{mis}}$ form the only memory in the MICE algorithm. Chapter 3 explained that imputed data can have a considerable amount of random noise, depending on the strength of the relations between the variables. Applications of MICE with lowly correlated data therefore inject a lot of noise into the system. Hence, the autocorrelation over t will be low, and convergence will be rapid, and in fact immediate if all variables are independent. Thus, the incorporation of noise into the imputed data has the pleasant side effect of speeding up convergence. Conversely, situations to watch out for may occur if:

- the correlations between the Y_js are high;
- the missing data rates are high; or
- constraints on parameters across different variables exist.

The first two conditions directly affect the amount of autocorrelation in the system. The latter condition becomes relevant for customized imputation models. We will see some examples in Section 5.5.2.

In the context of missing data imputation, our simulations have shown that unbiased estimates and appropriate coverage usually require no more than just five iterations. It is, however, important not to rely automatically on this result as some applications can require considerably more iterations.

4.5.6 Example of slow convergence

Consider a small simulation experiment with three variables: one complete covariate X and two incomplete variables Y_1 and Y_2. The data consist of draws from the multivariate normal distribution with correlations $\rho(X, Y_1) = \rho(X, Y_2) = 0.9$ and $\rho(Y_1, Y_2) = 0.7$. The variables are ordered as $[X, Y_1, Y_2]$. The complete pattern is $R_1 = (1, 1, 1)$. Missing data are randomly created in two patterns: $R_2 = (1, 0, 1)$ and $R_3 = (1, 1, 0)$. Variables Y_1 and Y_2 are jointly observed on $n_{(1,1,1)}$ complete cases. The following code defines the function to generate the incomplete data.

```
> generate <- function(n = c(1000, 4500, 4500, 0),
      cor = matrix(c(1, 0.9, 0.9, 0.9, 1, 0.7, 0.9,
          0.7, 1), nrow = 3)) {
      require(MASS)
      nt <- sum(n)
      cs <- cumsum(n)
      data <- mvrnorm(nt, mu = rep(0, 3), Sigma = cor)
```

```
        dimnames(data) <- list(1:nt, c("X", "Y1",
            "Y2"))
        if (n[2] > 0)
            data[(cs[1] + 1):cs[2], "Y1"] <- NA
        if (n[3] > 0)
            data[(cs[2] + 1):cs[3], "Y2"] <- NA
        if (n[4] > 0)
            data[(cs[3] + 1):cs[4], c("Y1", "Y2")] <- NA
        return(data)
    }
```

As an imputation model, we specified compatible linear regressions $Y_1 = \beta_{1,0} + \beta_{1,2}Y_2 + \beta_{1,3}X + \epsilon_1$ and $Y_2 = \beta_{2,0} + \beta_{2,1}Y_1 + \beta_{2,3}X + \epsilon_2$ to impute Y_1 and Y_2. The following code defines the function used for imputation.

```
> impute <- function(data, m = 5, method = "norm",
      print = FALSE, maxit = 10, ...) {
      statistic <- matrix(NA, nrow = maxit, ncol = m)
      for (iter in 1:maxit) {
          if (iter == 1)
              imp <- mice(data, m = m, method = method,
                  print = print, maxit = 1, ...)
          else imp <- mice.mids(imp, maxit = 1,
              print = print, ...)
          statistic[iter, ] <- unlist(with(imp,
              cor(Y1, Y2))$analyses)
      }
      return(list(imp = imp, statistic = statistic))
  }
```

The difficulty in this particular problem is that the correlation $\rho(Y_1, Y_2)$ under the conditional independence of Y_1 and Y_2 given X is equal to $0.9 \times 0.9 = 0.81$, whereas the true value equals 0.7. It is thus of interest to study how the correlation $\rho(Y_1, Y_2)$ develops over the iterations, but this is not a standard function in `mice()`. As an alternative, the `impute()` function repeatedly calls `mice.mids()` with `maxit = 1`, and calculates $\rho(Y_1, Y_2)$ after each iteration from the complete data.

The following code defines six scenarios where the number of complete cases is varied as $n_{(1,1,1)} \in \{1000, 500, 250, 100, 50, 0\}$, while holding the total sample size constant at $n = 10,000$. The proportion of complete rows thus varies between 10% and 0%.

```
> simulate <- function(ns = matrix(c(1000, 500,
      250, 100, 50, 0, rep(c(4500, 4750, 4875, 4950,
          4975, 5000), 2), rep(0, 6)), nrow = 6),
      m = 5, maxit = 10, seed = 1, ...) {
      if (!missing(seed))
```

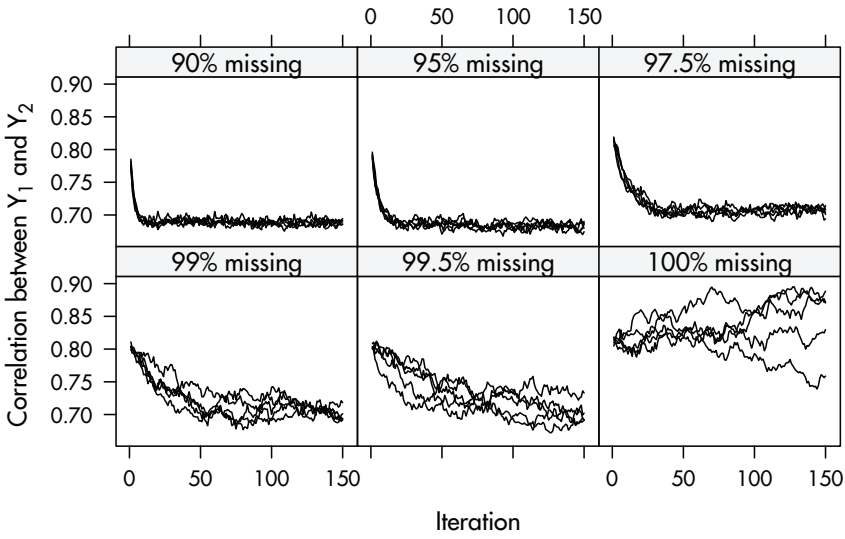


Figure 4.3: Correlation between Y_1 and Y_2 in the imputed data per iteration in five independent runs of the MICE algorithm for six levels of missing data. The true value is 0.7. The figure illustrates that convergence can be slow for high percentages of missing data.

```
    set.seed(seed)
  s <- cbind(rep(1:nrow(ns), each = maxit *
    m), apply(ns, 2, rep, each = maxit * m),
    rep(1:maxit, each = m), 1:m, NA)
  colnames(s) <- c("k", "n111", "n101", "n110",
    "n100", "iteration", "m", "rY1Y2")
  for (k in 1:nrow(ns)) {
    data <- generate(ns[k, ], ...)
    r <- impute(data, m = m, maxit = maxit,
      ...)
    s[s[, "k"] == k, "rY1Y2"] <- t(r$statistic)
  }
  return(data.frame(s))
}
```

The `simulate()` function code collects the correlations $\rho(Y_1, Y_2)$ per iteration in the data frame `s`. Now call the function with

```
> slow.demo <- simulate(maxit = 150, seed = 62771)
```

Figure 4.3 shows the development of $\rho(Y_1, Y_2)$ calculated on the completed data after every iteration of the MICE algorithm. At iteration 1, $\rho(Y_1, Y_2)$ is approximately 0.81, the value expected under independence of Y_1 and Y_2, conditional on X. The influence of the complete records with both Y_1 and Y_2 observed percolates into the imputations, so that the chains slowly move into the direction of the population value of 0.7. The speed of convergence heavily depends on the number of missing cases. For 90–95% missing data, the streams are essentially flat after about 15–20 iterations. As the percentage of missing data increases, more and more iterations are needed before the true correlation of 0.7 trickles through. In the extreme cases with 100% missing data, the correlation $\rho(Y_1, Y_2)$ cannot be estimated due to lack of information in the data. In this case, the different streams do not converge at all, and wander widely within the Cauchy–Schwarz bounds (0.6 to 1.0 here). But even here we could argue that the sampler has essentially converged. We could stop at iteration 200 and take the imputations from there. From a Bayesian perspective, this still would yield an essentially correct inference about $\rho(Y_1, Y_2)$, being that it could be anywhere within the Cauchy–Schwarz bounds. So even in this pathological case with 100% missing data, the results look sensible as long as we account for the wide variability.

The lesson we can learn from this simulation is that we should be careful about convergence in missing data problems with high correlations and high missing data rates. At the same time, observe that we really have to push the MICE algorithm to its limits to see the effect. Over 99% of real data will have lower correlations and lower missing data rates. Of course, it never hurts to do a couple of extra iterations, but my experience is that good results can often be obtained with a small number of iterations.

4.6 FCS and JM

4.6.1 Relations between FCS and JM

FCS is related to JM in some special cases. If $P(X, Y)$ has a multivariate normal model distribution, then all conditional densities are linear regressions with a constant normal error variance. So, if $P(X, Y)$ is multivariate normal then $P(Y_j|X, Y_{-j})$ follows a linear regression model. The reverse is also true: If the imputation models $P(Y_j|X, Y_{-j})$ are all linear with constant normal error variance, then the joint distribution will be multivariate normal. See Arnold et al. (1999, p. 186) for a description of the precise conditions. Thus, imputation by FCS using all linear regressions is identical to imputation under the multivariate normal model.

Another special case occurs for binary variables with only two-way interactions in the log-linear model. For example, in the case $p = 3$ suppose that

$Y_1, \ldots, Y_3$ are modeled by the log-linear model that has the three-way interaction term set to zero. It is known that the corresponding conditional distribution $P(Y_1|Y_2, Y_3)$ is the logistic regression model $\log(P(Y_1)/1 - P(Y_1)) = \beta_0 + \beta_2 Y_2 + \beta_3 Y_3$ (Goodman, 1970). Analogous definitions exist for $P(Y_2|Y_1, Y_3)$ and $P(Y_3|Y_1, Y_2)$. This means that if we use logistic regressions for Y_1, Y_2 and Y_3, we are effectively imputing under the multivariate "no three-way interaction" log-linear model.

4.6.2 Comparison

FCS cannot use computational shortcuts like the sweep operator, so the calculations per iterations are more intensive than under JM. Also, JM has better theoretical underpinnings.

On the other hand, FCS allows tremendous flexibility in creating multivariate models. One can easily specify models that are outside any known standard multivariate density $P(X, Y, R|\theta)$. FCS can use specialized imputation methods that are difficult to formulate as a part of a multivariate density $P(X, Y, R|\theta)$. Imputation methods that preserve unique features in the data, e.g., bounds, skip patterns, interactions, bracketed responses and so on can be incorporated. It is possible to maintain constraints between different variables in order to avoid logical inconsistencies in the imputed data that would be difficult to do as part of a multivariate density $P(X, Y, R|\theta)$.

4.6.3 Illustration

The Fourth Dutch Growth Study by Fredriks et al. (2000a) collected data on 14500 Dutch children between 0 and 21 years. The development of secondary pubertal characteristics was measured by the so-called Tanner stages, which divides the continuous process of maturation into discrete stages for the ages between 8 and 21 years. Pubertal stages of boys are defined for genital development (`gen`: five ordered stages G1–G5), pubic hair development (`phb`: six ordered stages P1–P6) and testicular volume (`tv`: 1–25 ml).

We analyze the subsample of 424 boys in the age range 8–21 years. There were 180 boys (42%) for which scores for genital development were missing. The missingness was strongly related to age, rising from about 20% at ages 9–11 years to 60% missing data at ages 17–20 years.

The data consist of three complete covariates: age (`age`), height (`hgt`) and weight (`wgt`), and three incomplete outcomes measuring maturation. First, the data are imputed $m = 10$ times under the multivariate normal model for all variables as follows:

```
> select <- with(boys, age >= 8 & age <= 21)
> djm <- boys[select, -4]
> djm$gen <- as.integer(djm$gen)
> djm$phb <- as.integer(djm$phb)
```

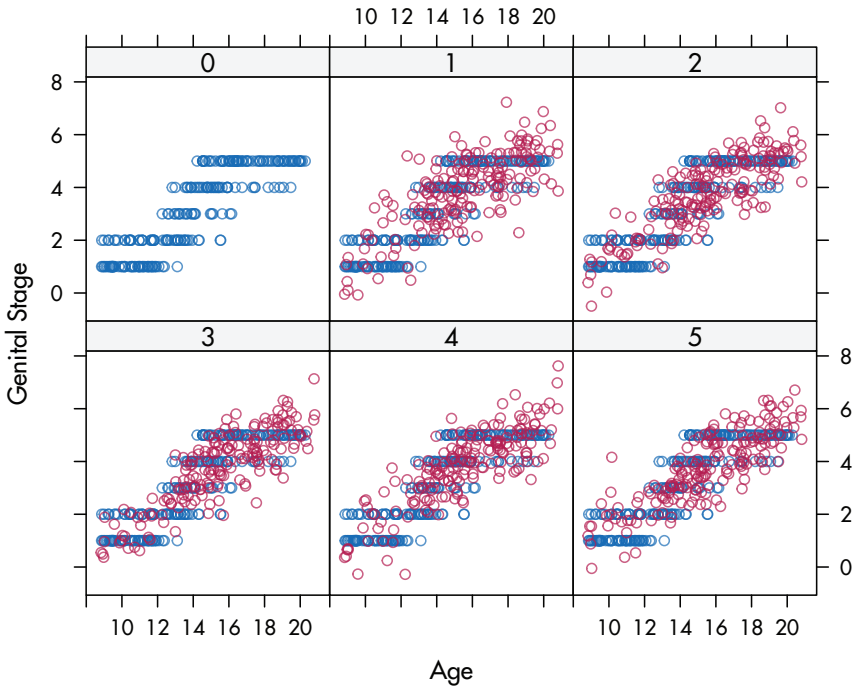


Figure 4.4: Joint modeling: Imputed data for genital development (Tanner stages G1–G5) under the multivariate normal model. The panels are labeled by the imputation numbers 0–5, where 0 is the observed data and 1–5 are five multiply imputed datasets.

```
> djm$reg <- as.integer(djm$reg)
> jm.10 <- mice(djm, method = "norm", seed = 93005,
      m = 10)
```

Figure 4.4 plots the results of the first five imputations. It was created by the following statement:

```
> xyplot(jm.10, gen ~ age | .imp, subset = as.integer(.imp) <
      7, ylab = "Genital stage")
```

The figure portrays how genital development depends on age for both the observed and imputed data. The spread of the synthetic values in Figure 4.4 is larger than the observed data range. The observed data are categorical while the synthetic data vary continuously. Note that there are some negative values in the imputations. If we are to do categorical data analysis on the imputed data, we need some form of rounding to make the synthetic values comparable with the observed values.

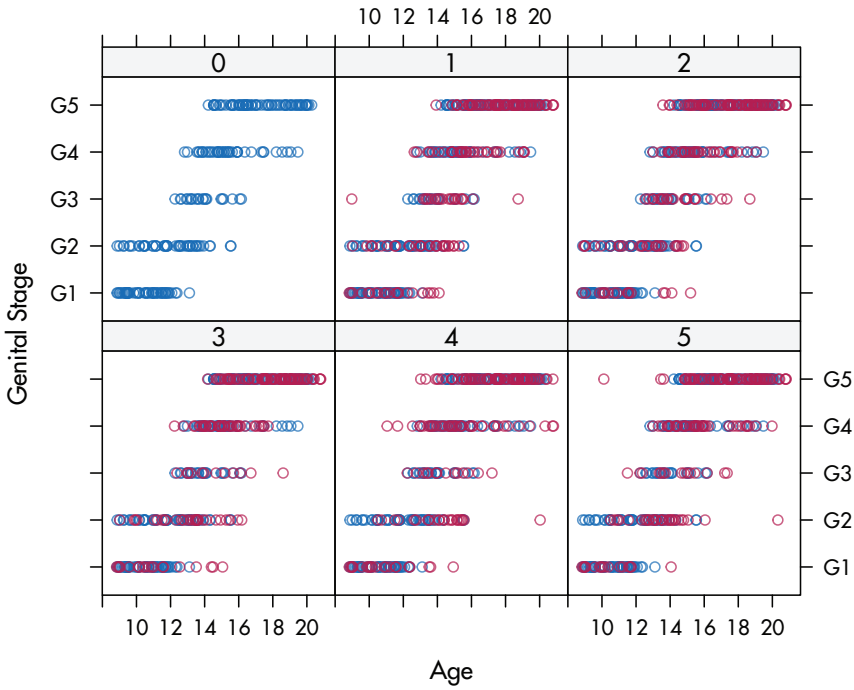


Figure 4.5: Fully conditional specification: Imputed data of genital development (Tanner stages G1–G5) under the proportional odds model.

Multiple imputations by FCS for a combination of predictive mean matching (for the continuous variables) and the proportional odds model (for the ordered variables `gen` and `phb`) can be calculated and plotted as

```
> fcs.10 <- mice(dfcs, seed = 81420, m = 10)
> xyplot(fcs.10, gen ~ age | .imp, subset = as.integer(.imp) <
      7, ylab = "Genital stage")
```

The imputations in Figure 4.5 differ markedly from those in Figure 4.4. The proportonal odds model yields imputations that are categorical, and hence no rounding is needed.

The complete data model describes the probability of achieving each Tanner stage as a nonlinear function of age according to the model proposed in Van Buuren and Ooms (2009). The calculations are done with `gamlss` (Stasinopoulos and Rigby, 2007). Under the assumption of ignorability, analysis of the complete cases will not be biased, so the complete case analysis provides a handle to the appropriate solution. The blue lines in Figure 4.6

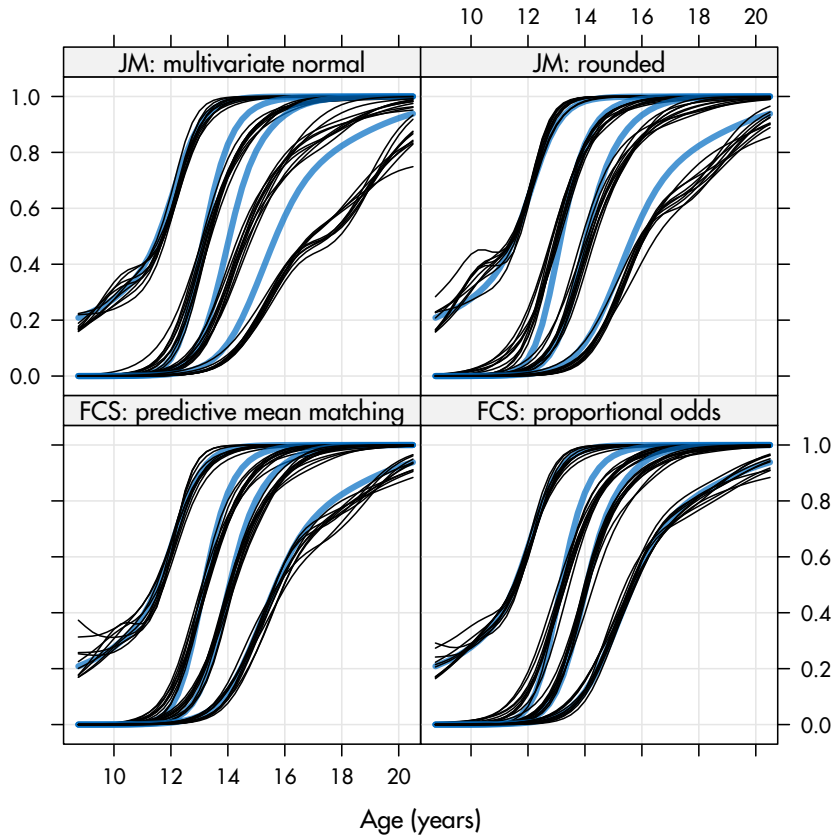


Figure 4.6: Probability of achieving stages G2–G5 of genital developmental by age (in years) under four imputation methods ($m = 10$).

indicate the model fitted on the complete cases, whereas the thin black lines correspond to the analyses of the 10 imputed datasets.

The different panels of Figure 4.6 corresponds to different imputation methods. The panel labeled *JM: multivariate normal* contains the model fitted to the unprocessed imputed data produced under the multivariate normal model. There is a large discrepancy between the complete case analysis and the models fitted to the imputed data, especially for the older boys. The fit improves in the panel labeled *JM: rounded*, where imputed data are rounded to the nearest category. There is considerable misfit, and the behavior of the imputed data around the age of 10 years is a bit curious. The panel labeled *FCS: predictive mean matching* applied Algorithm 3.3 within the MICE algorithm. Though this technique improves upon the previous two methods, some

discrepancies for the older boys remain. The panel labeled *FCS: proportional odds* displays the results after applying the method for ordered categorical data as discussed in Section 3.5. The imputed data essentially agree with the complete case analysis, perhaps apart from some minor deviations around the probability level of 0.9.

Figure 4.6 shows clear differences between FCS and JM when data are categorical. Although rounding may provide reasonable results in particular datasets, it seems that it does more harm than good here. There are many ways to round, rounding may require unrealistic assumptions and it will attenuate correlations. Horton et al. (2003), Ake (2005) and Allison (2005) recommend against rounding when data are categorical. See Section 4.4.3. Horton et al. (2003) expected that bias problems of rounding would taper off if variables have more than two categories, but the analysis in this section suggests that JM may also be biased for categorical data with more than two categories. Even though it may sound a bit trivial, my recommendation is: Impute categorical data by methods for categorical data.

4.7 Conclusion

Multivariate missing data lead to analytic problems caused by mutual dependencies between incomplete variables. The missing data pattern provides important information for the imputation model. The influx and outflux measures are useful to sift out variables that cannot contribute to the imputations. For general missing data patterns, both JM and FCS approaches can be used to impute multivariate missing data. JM is the model of choice if the data conform to the modeling assumptions because it has better theoretical properties. The FCS approach is much more flexible and allows for imputations close to the data. Lee and Carlin (2010) provide a comparison between both perspectives.

4.8 Exercises

1. *MAR (continued)*. Repeat Exercise 3.1 for a multivariate missing data mechanism.

2. *Convergence*. Figure 4.3 shows that convergence can take longer for very high amounts of missing data. This exercise studies an even more extreme situation.

(a) The default argument `ns` of the `simulate()` function in Section 4.5.6 defines six scenarios with different missing data patterns. Define a 6×4 matrix `ns2`, where patterns R_2 and R_3 are replaced by pattern $R_4 = (1, 0, 0)$. How many more missing values are there in each scenario?

(b) For the new scenarios, do you expect convergence to be slower or faster? Explain.

(c) Change the scenario in which all data in Y_1 and Y_2 are missing so that there are 20 complete cases. Then run

```
> slow2 <- simulate(ns = ns2, maxit = 50, seed = 62771)
```

and create a figure similar to Figure 4.3.

(d) Compare your figure with Figure 4.3. Are there any major differences? If so, which?

(e) Did the figure confirm your idea about convergence speed you had formulated in (b)?

(f) How would you explain the behavior of the trace lines?

3. *Binary data.* Perform the simulations of Section 4.5.6 with binary Y_1 and Y_2. Use the odds ratio instead of the correlation to measure the association between Y_1 and Y_2. Does the same conclusion hold?

Chapter 5

Imputation in practice

Chapters 3 and 4 describe methods to generate multiple imputations. The application of these techniques in practice should be done with appropriate care. This chapter focuses on practical issues that surround the methodology. This chapter assumes that multiple imputations are created by means of the MICE algorithm, as described in Section 4.5.2.

This chapter relies on the `R` package `mice` to implement the techniques. Many of these have equivalents in other software, in particular the `R` packages `Hmisc`, `mi`, `miP` and `VIM`, the `Stata` commands `mi` and `ice`, and the SAS callable software application `IVEware`. See Appendix A for an overview of software for multiple imputation.

5.1 Overview of modeling choices

The specification of the imputation model is the most challenging step in multiple imputation. The imputation model should

- account for the process that created the missing data,
- preserve the relations in the data, and
- preserve the uncertainty about these relations.

The idea is that adherence to these principles will yield proper imputations (cf. Section 2.3.3), and thus result in valid statistical inferences. What are the choices that we need to make, and in what order? Van Buuren and Groothuis-Oudshoorn (2011) list the following seven choices:

1. First, we should decide whether the MAR assumption is plausible. See Sections 1.2 and 2.2.4 for an introduction to MAR and MNAR. Chained equations can handle both MAR and MNAR. Multiple imputation under MNAR requires additional modeling assumptions that influence the generated imputations. There are many ways to do this. Section 3.9 described one way to do so within the FCS framework. Section 5.2 deals with this issue in more detail.

2. The second choice refers to the form of the imputation model. The form encompasses both the structural part and the assumed error distribution. In FCS the form needs to be specified for each incomplete column in the data. The choice will be steered by the scale of the variable to be imputed, and preferably incorporates knowledge about the relation between the variables. Chapter 3 described many different methods for creating univariate imputations.

3. A third choice concerns the set of variables to include as predictors in the imputation model. The general advice is to include as many relevant variables as possible, including their interactions (Collins et al., 2001). This may, however, lead to unwieldy model specifications. Section 5.3 describes the facilities within the `mice()` function for setting the predictor matrix.

4. The fourth choice is whether we should impute variables that are functions of other (incomplete) variables. Many datasets contain derived variables, sum scores, interaction variables, ratios and so on. It can be useful to incorporate the transformed variables into the multiple imputation algorithm. Section 5.4 describes methods that we can use to incorporate such additional knowledge about the data.

5. The fifth choice concerns the order in which variables should be imputed. The visit sequence may affect the convergence of the algorithm and the synchronization between derived variables. Section 5.5.1 discusses relevant options.

6. The sixth choice concerns the setup of the starting imputations and the number of iterations. The convergence of the MICE algorithm can be monitored in many ways. Section 5.5.2 outlines some techniques that assist in this task.

7. The seventh choice is m, the number of multiply imputed datasets. Setting m too low may result in large simulation error and statistical inefficiency, especially if the fraction of missing information is high. Section 2.7 provided guidelines for setting m.

Please realize that these choices are always needed. Imputation software needs to make default choices. These choices are intended to be useful across a wide range of applications. However, the default choices are not necessarily the best for the data at hand. There is simply no magical setting that always works, so often some tailoring is needed. Section 5.6 highlights some diagnostic tools that aid in determining the choices.

5.2 Ignorable or nonignorable?

Recall from Section 2.2.6 that the assumption of ignorability is essentially the belief that the available data are sufficient to correct the missing data. There are two main strategies that we might pursue if the response mechanism is nonignorable:

- Expand the data in the imputation model in the hope of making the missing data mechanism closer to MAR, or
- Formulate and fit a nonignorable imputation model and perform sensitivity analysis on the critical parameters.

Collins et al. (2001) remarked that it is a "safe bet" there will be lurking variables Z that are correlated both with the variables of interest Y and with the missingness of Y. The important question is, however, whether these correlations are strong enough to produce substantial bias if no measures are taken. Collins et al. (2001) performed simulations that provided some answers in the case of linear regression. If the missing data rate did not exceed 25% and if the correlation between the Z and Y was 0.4, omitting Z from the imputation model had a negligible effect. For more extreme situations, with 50% missing data and/or a correlation of 0.9, the effect depended strongly on the form of the missing data mechanism. When the probability to be missing was linear in Z (like MARRIGHT in Section 3.2.4), then omitting Z from the imputation model only affected the intercept, whereas the regression weights and variance estimates were unaffected. When more missing data were created in the extremes (like MARTAIL), the reverse occurred: omitting Z affected the regression coefficients and variance estimates, but the intercept was unbiased with the correct confidence interval. In summary, all estimates under multiple imputation were remarkably robust against MNAR in many instances. Beyond a correlation of 0.4 or a missing data rate over 25% the form of the missing data mechanism determines which parameters are affected.

Based on these results, we suggest the following guidelines. The MAR assumption is often a suitable starting point. If the MAR assumption is suspect for the data at hand, a next step is to find additional data that are strongly predictive of the missingness, and include these into the imputation model. If all possibilities for such data are exhausted and if the assumption is still suspect, perform a concise simulation study as in Collins et al. (2001) customized for the problem at hand with the goal of finding out how extreme the MNAR mechanism needs to be to influence the parameters of scientific interest. Finally, use a nonignorable imputation model (cf. Section 3.9) to correct the direction of imputations created under MAR. Vary the most critical parameters, and study their influence on the final inferences. Section 7.2 contains an example of how this can be done in practice.

Table 5.1: Built-in univariate imputation techniques in the `mice` package.

Method	Description	Scale Type
pmm	Predictive mean matching	Numeric*
norm	Bayesian linear regression	Numeric
norm.predict	Predicted value	Numeric
norm.nob	Stochastic regression	Numeric
norm.boot	Normal imputation with bootstrap	Numeric
mean	Unconditional mean imputation	Numeric
2L.norm	Multilevel normal model	Numeric
logreg	Logistic regression	Binary*
logreg.boot	Logistic regression with bootstrap	Binary
polyreg	Multinomial logit model	Nominal*
lda	Discriminant analysis	Nominal
polr	Ordered logit model	Ordinal*
sample	Simple random sample	Any

* = default for scale type

5.3 Model form and predictors

5.3.1 Model form

The MICE algorithm requires a specification of a univariate imputation method separately for each incomplete variable. Chapter 3 discussed many possible methods. The measurement level largely determines the form of the univariate imputation model. The `mice()` function distinguishes numerical, binary, ordered and unordered categorical data, and sets the defaults accordingly.

Table 5.1 lists the built-in univariate imputation method in the `mice` package. The defaults have been chosen to work well in a wide variety of situations, but in particular cases different methods may be better. For example, if it is known that the variable is close to normally distributed, using `norm` instead of the default `pmm` may be more efficient. For large datasets where sampling variance is not an issue, it could be useful to select `norm.nob`, which does not draw regression parameters, and is thus simpler and faster. The `norm.boot` method is a fast non-Bayesian alternative for `norm`. The `norm` methods are an alternative to `pmm` in cases where `pmm` does not work well, e.g., when insufficient nearby donors can be found.

The `mean` method is included for completeness and should not be generally used. The `2L.norm` is meant to impute two-level normal data. For very sparse categorical data, it may be better to use `pmm` instead of `logreg`, `polr` or `polyreg`. Method `logreg.boot` is a version of `logreg` that uses the bootstrap to emulate sampling variance. Method `lda` is generally inferior to `polyreg`

(Brand, 1999), and should be used only as a backup when all else fails. Finally, `sample` is a quick method for creating starting imputations without the need for covariates.

5.3.2 Predictors

A useful feature of the `mice()` function is the ability to specify the set of predictors to be used for each incomplete variable. The basic specification is made through the `predictorMatrix` argument, which is a square matrix of size `ncol(data)` containing 0/1 data. Each row in `predictorMatrix` identifies which predictors are to be used for the variable in the row name. If `diagnostics=T` (the default), then `mice()` returns a `mids` object containing a `predictorMatrix` entry. For example, type

```
> imp <- mice(nhanes, print = FALSE)
> imp$predictorMatrix
    age bmi hyp chl
age   0   0   0   0
bmi   1   0   1   1
hyp   1   1   0   1
chl   1   1   1   0
```

The rows correspond to incomplete target variables, in the sequence as they appear in the data. A value of `1` indicates that the column variable is a predictor to impute the target (row) variable, and a `0` means that it is not used. Thus, in the above example, `bmi` is predicted from `age`, `hyp` and `chl`. Note that the diagonal is `0` since a variable cannot predict itself. Since `age` contains no missing data, `mice()` silently sets all values in the row to `0`. The default setting of the `predictorMatrix` specifies that every variable predicts all others.

Conditioning on all other data is often reasonable for small to medium datasets, containing up to, say, 20–30 variables, without derived variables, interactions effects and other complexities. As a general rule, using every bit of available information yields multiple imputations that have minimal bias and maximal efficiency (Meng, 1994; Collins et al., 2001). It is often beneficial to choose as large a number of predictors as possible. Including as many predictors as possible tends to make the MAR assumption more plausible, thus reducing the need to make special adjustments for MNAR mechanisms (Schafer, 1997).

For datasets containing hundreds or thousands of variables, using all predictors may not be feasible (because of multicollinearity and computational problems) to include all these variables. It is also not necessary. In my experience, the increase in explained variance in linear regression is typically negligible after the best, say, 15 variables have been included. For imputation purposes, it is expedient to select a suitable subset of data that contains no

more than 15 to 25 variables. Van Buuren et al. (1999) provide the following strategy for selecting predictor variables from a large database:

1. Include all variables that appear in the complete data model, i.e., the model that will be applied to the data after imputation, including the outcome (Little, 1992; Moons et al., 2006). Failure to do so may bias the complete data analysis, especially if the complete data model contains strong predictive relations. Note that this step is somewhat counter-intuitive, as it may seem that imputation would artificially strengthen the relations of the complete data model, which would be clearly undesirable. If done properly however, this is not the case. On the contrary, not including the complete data model variables will tend to bias the results toward zero. Note that interactions of scientific interest also need to be included in the imputation model.

2. In addition, include the variables that are related to the nonresponse. Factors that are known to have influenced the occurrence of missing data (stratification, reasons for nonresponse) are to be included on substantive grounds. Other variables of interest are those for which the distributions differ between the response and nonresponse groups. These can be found by inspecting their correlations with the response indicator of the variable to be imputed. If the magnitude of this correlation exceeds a certain level, then the variable should be included.

3. In addition, include variables that explain a considerable amount of variance. Such predictors help reduce the uncertainty of the imputations. They are basically identified by their correlation with the target variable.

4. Remove from the variables selected in steps 2 and 3 those variables that have too many missing values within the subgroup of incomplete cases. A simple indicator is the percentage of observed cases within this subgroup, the percentage of usable cases (cf. Section 4.1.2).

Most predictors used for imputation are incomplete themselves. In principle, one could apply the above modeling steps for each incomplete predictor in turn, but this may lead to a cascade of auxiliary imputation problems. In doing so, one runs the risk that every variable needs to be included after all.

In practice, there is often a small set of key variables, for which imputations are needed, which suggests that steps 1 through 4 are to be performed for key variables only. This was the approach taken in Van Buuren and Groothuis-Oudshoorn (1999), but it may miss important predictors of predictors. A safer and more efficient, though more laborious, strategy is to perform the modeling steps also for the predictors of predictors of key variables. This is done in Groothuis-Oudshoorn et al. (1999). I expect that it is rarely necessary to go beyond predictors of predictors. At the terminal node, we can apply a simple

method like `mice.impute.sample()` that does not need any predictors for itself.

The `mice` package contains several tools that aid in automatic predictor selection. The `quickpred()` function is a quick way to define the predictor matrix using the strategy outlined above. The `flux()` was described in Section 4.1.3. The `mice()` function detects multicollinearity, and solves the problem by removing one or more predictors for the matrix. Each removal is noted in the `loggedEvents` element of the `mids` object. For example,

```
> imp <- mice(cbind(nhanes, chl2 = 2 * nhanes$chl),
      print = FALSE)
> imp$loggedEvents
  it im co dep      meth  out
1  0  0  0     collinear chl2
```

informs us that the duplicate variable `chl2` was removed before iteration. The algorithm also detects multicollinearity during iterations. Another measure to control the algorithm is the `ridge` parameter, denoted by κ in Algorithm 3.1, 3.2 and 3.3. The `ridge` parameter is specified as an argument to `mice()`. Setting `ridge=0.001` or `ridge=0.01` makes the algorithm more robust at the expense of bias.

There is some room for improvement by providing automatic variable selection that selects the, say, 20 best variables from all data. It is, however, difficult to fully automate model building since selection according to steps 1 and 2 cannot be mechanical.

Care is needed if the imputation model contains derived variables, like transformations, recodes or interaction terms. The default specification may lock up the MICE algorithm or produce erratic imputations. As a general, rule, feedback between different versions of the same variable should be prevented. The next section describes a number of techniques are useful in various situations.

5.4 Derived variables

5.4.1 Ratio of two variables

In practice, there is often extra knowledge about the data that is not modeled explicitly. For example, consider the weight/height ratio `whr`, defined as `wgt`/`hgt` (kg/m). If any one of the triplet (`hgt`, `wgt`, `whr`) is missing, then the missing value can be calculated with certainty by a simple deterministic rule. Unless we specify otherwise, the imputation model is unaware of the relation between the three variables, and will produce imputations that are inconsistent with the rule. Inconsistent imputations are clearly undesirable since they yield

combinations of data values that are impossible had the data been observed. Including knowledge about derived data in the imputation model prevents imputations from being inconsistent. Knowledge about the derived data can take many forms, and includes data transformations, interactions, sum scores, recoded versions, range restrictions, if-then relations and polynomials.

The easiest way to deal with the problem is to leave any derived data outside the imputation process. For example, we may impute any missing height and weight data, and append `whr` to the imputed data afterward. It is simple to do that in `mice` by

```
> imp1 <- mice(boys)
> long <- complete(imp1, "long", inc = TRUE)
> long$whr <- with(long, wgt/(hgt/100))
> imp2 <- long2mids(long)
```

Another possibility is to create `whr` before imputation, and impute `whr` as "just another variable," an approach known as JAV (White et al., 2011b):

```
> boys$whr <- boys$wgt/(boys$hgt/100)
> imp.jav <- mice(boys, m = 1, seed = 32093, maxit = 10)
```

Although JAV may yield valid statistical inferences in particular cases, it may easily produce absurd combinations of imputed values. For example, consider the scatterplot of `hgt` and `whr`. The leftmost panel in Figure 5.1 alerts us that something is wrong. The reason for this is twofold. First, we did not specify that `whr` is actually a function of `hgt` and `wgt`. Second, adding `whr` to the `boys` data introduced linear dependencies. The `mice()` function handled this by removing one of the predictors for each imputation model. We can see which variables are removed by inspecting the logged events:

```
> imp$loggedEvents
   it im co dep meth out
1   1  1 10 whr  pmm wgt
2   2  1  6 gen polr whr
3   2  1  7 phb polr whr
...
```

which informs us that `whr` or `wgt` are removed from some models. Also, observe that JAV does not preserve the known deterministic relations in the data.

A solution for this problem is a special mechanism, called *passive imputation*. Passive imputation maintains the consistency among different transformations of the same data. The method can be used to ensure that the transform always depends on the most recently generated imputations in the original untransformed data.

In `mice` passive imputation is invoked by specifying "~" as the first character of the imputation method. This provides a simple method for specifying

a large variety of dependencies among the variables, such as transformed variables, recodes, interactions, sum scores and so on.

In the above example, we invoke passive imputation by

```
> ini <- mice(boys, m = 1, maxit = 0)
> meth <- ini$meth
> meth["whr"] <- "~I(wgt/(hgt/100))"
> meth["bmi"] <- "~I(wgt/(hgt/100)^2)"
> pred <- ini$pred
> pred[c("wgt", "hgt", "bmi"), "whr"] <- 0
> pred[c("wgt", "hgt", "whr"), "bmi"] <- 0
> pred
     age hgt wgt bmi hc gen phb tv reg whr
age    0   0   0   0  0   0   0  0   0   0
hgt    1   0   1   0  1   1   1  1   1   0
wgt    1   1   0   0  1   1   1  1   1   0
bmi    1   1   1   0  1   1   1  1   1   0
hc     1   1   1   1  0   1   1  1   1   1
gen    1   1   1   1  1   0   1  1   1   1
phb    1   1   1   1  1   1   0  1   1   1
tv     1   1   1   1  1   1   1  0   1   1
reg    1   1   1   1  1   1   1  1   0   1
whr    1   1   1   0  1   1   1  1   1   0


> imp.pas <- mice(boys, m = 1, meth = meth, pred = pred,
      seed = 32093, maxit = 10)
```

The `I()` operator in the `meth` definitions instructs `R` to interpret the argument as literal. So `I(wgt/(hgt/100))` calculates `whr` by dividing `wgt` by `hgt` (in meters). A similar statement is given for `bmi`. The imputed values for the variables `whr` and `bmi` are thus derived from `hgt` and `wgt` according to the stated methods, and hence are consistent. The changes to the default predictor matrix are needed to break any feedback loops between the derived variables and their originals. It is important to do this. Using `whr` to impute `hgt` or `wgt` would result in absurd imputations and problematic convergence.

The middle panel in Figure 5.1 shows that passive imputation represents an improvement over JAV. Though some of the imputed data points appear extreme, the values are generally similar to the real data. Moreover, the values adhere to the derived rules. It is possible to create slightly better imputations by preventing that `bmi`, `whr` and the pair (`hgt`, `wgt`) are simultaneous predictors for any variable as follows:

```
> pred[c("wgt", "hgt", "hc", "reg"), "bmi"] <- 0
> pred[c("gen", "phb", "tv"), c("hgt", "wgt", "hc")] <- 0
> pred[, "whr"] <- 0
```

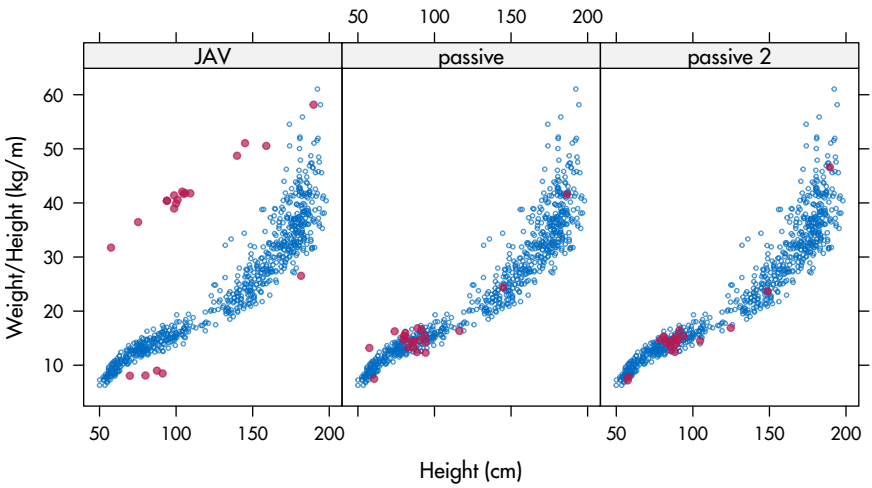


Figure 5.1: Three different imputation models to impute weight/height ratio (`whr`). The relation between `whr` and height (`hgt`) is not respected under "just another variable" (JAV). Both passive methods yield imputations that are close to the observed data. "Passive 2" does not allow for models in which `whr` and `bmi` are simultaneous predictors.

```
> imp.pas2 <- mice(boys, m = 1, meth = meth, pred = pred,
      seed = 32093, maxit = 10)
```

Passive imputation overrules the selection of variables in the `predictorMatrix` argument. Thus, in the above case, we might as well have set `pred["bmi",] <- 0` or `pred["whr",] <- 0` and obtained identical results.

The usefulness of JAV depends somewhat on the nature of the complete-data model. When the primary interest focuses exclusively on regression weights and when the data are MCAR, JAV can yield valid inferences (Von Hippel, 2009). JAV, however, ignores any deterministic relations in the data, potentially leading to impossible combinations. When the ensuing analyses require consistency of the data, JAV cannot be used. In such cases, passive imputation can help obtain unbiased estimates from imputed data that are consistent.

5.4.2 Sum scores

The sum score is undefined if one of the variables to be added is missing. We can use sum scores of imputed variables within the MICE algorithm to economize on the number of predictors. For example, suppose we create a summary maturation score of the pubertal measurements `gen`, `phb` and `tv`, and use that score to impute the other variables instead of the three original pu-

bertal measurements. This technique is also useful for calculating scale scores from imputed questionnaire items (Van Buuren, 2010). It is easy to specify such models in `mice`. See Van Buuren and Groothuis-Oudshoorn (2011) for examples.

5.4.3 Interaction terms

The standard MICE algorithm only models main effects. Sometimes the interaction between variables is of scientific interest. For example, in a longitudinal study we could be interested in assessing whether the rate of change differs between two treatment groups, in other words, the treatment-by-group interaction. The standard algorithm does not take interactions into account, so the interactions of interest should be added to the imputation model.

The usual type of interactions between two continuous variables is to subtract the mean and take the product. The following code imputes the `boys` data with an interaction term of `wgt` and `hc` in the imputation model:

```
> expr <- expression((wgt - 40) * (hc - 50))
> boys$wgt.hc <- with(boys, eval(expr))
> ini <- mice(boys, max = 0)
> meth <- ini$meth
> meth["wgt.hc"] <- paste("~I(", expr, ")", sep = "")
> meth["bmi"] <- ""
> pred <- ini$pred
> pred[c("wgt", "hc"), "wgt.hc"] <- 0
> imp.int <- mice(boys, m = 1, maxit = 10, meth = meth,
      pred = pred, seed = 62587)
```

Figure 5.2 illustrates that the scatterplots of the real and synthetic values are similar. Furthermore, the imputations adhere to the stated recipe `(wgt-40)*(hc-50)`. Interactions involving categorical variables can be done in similar ways (Van Buuren and Groothuis-Oudshoorn, 2011).

Another way to deal with interaction is to impute the data in separate groups. One way of doing this in `mice` is to split the dataset into two or more parts, run `mice()` on each part and combine the imputed datasets with `rbind()`.

5.4.4 Conditional imputation

In some cases it makes sense to restrict the imputations, possibly conditional on other data. The method in Section 1.3.5 produced negative values for the positive-valued variable `Ozone`. One way of dealing with this mismatch between the imputed and observed values is to censor the values at some specified minimum or maximum value. The `mice()` function has an argument called `post` that takes a vector of strings of `R` commands. These commands are parsed and evaluated just after the univariate imputation function returns,

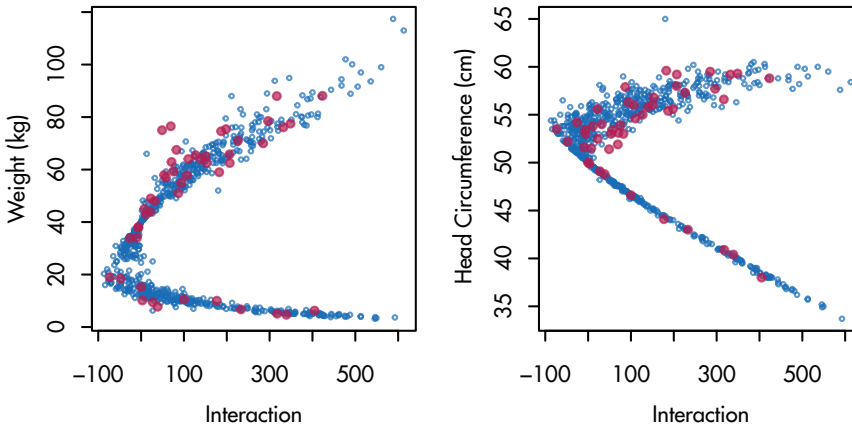


Figure 5.2: The relation between the interaction term `wgt.hc` (on the horizontal axes) and its components `wgt` and `hc` (on the vertical axes).

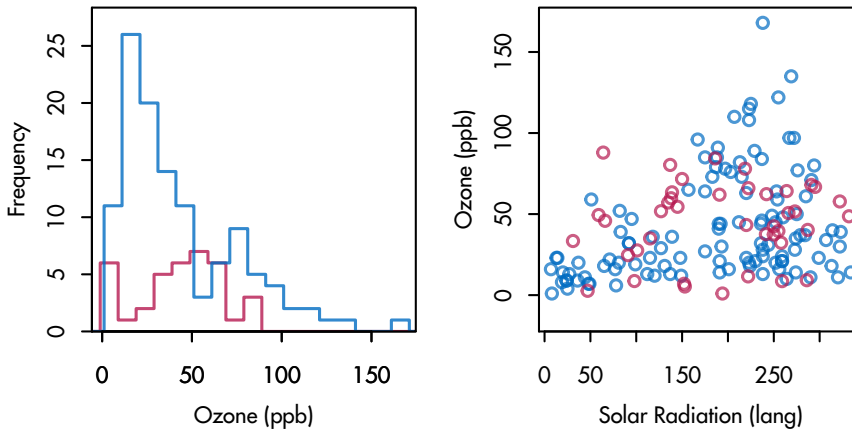


Figure 5.3: Stochastic regression imputation of `Ozone`, where the imputed values are restricted to the range 1–200. Compare to Figure 1.3.

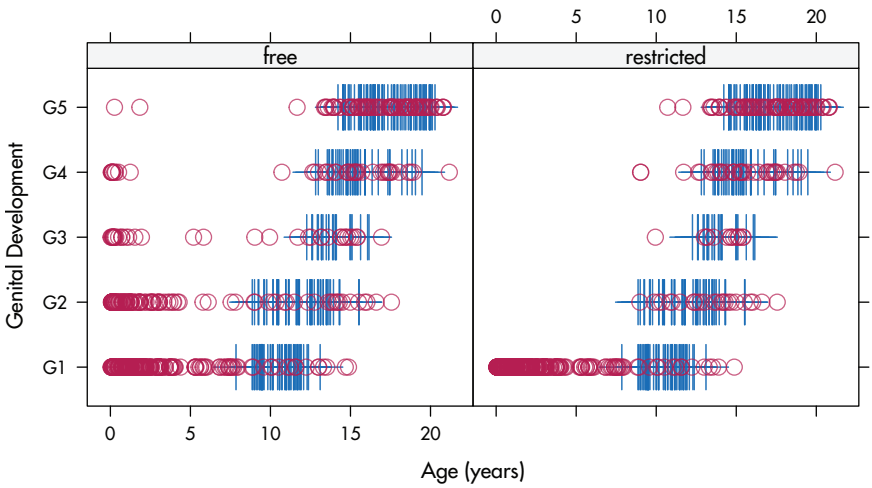


Figure 5.4: Genital development of Dutch boys by age. The "free" solution does not constrain the imputations, whereas the "restricted" solution requires all imputations below the age of 8 years to be at the lowest category.

and thus provide a way to post-process the imputed values. Note that `post` only affects the synthetic values, and leaves the observed data untouched. The `squeeze()` function in `mice` replaces values beyond the specified bounds by the minimum and maximal scale values. One way to ensure positive imputations for `Ozone` under stochastic regression imputation is

```
> ini <- mice(airquality[, 1:2], maxit = 0)
> post <- ini$post
> post["Ozone"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],
      c(1,200))"
> imp <- mice(airquality[, 1:2], method = "norm.nob",
      m = 1, maxit = 1, seed = 1, post = post)
```

Compare Figure 5.3 to Figure 1.3. The negative ozone value of −18.8 has now been replaced by a value of 1. The previous syntax of the `post` argument is a bit cumbersome. The same result can be achieved by neater code:

```
> post["Ozone"] <- "ifdo(c(Ozone<1, Ozone>200), c(1, 200))"
```

The `ifdo()` function is a convenient way to create conditional imputes. For example, in the `boys` data puberty is measured only for boys older than 8 years. Before this age it is unlikely that puberty has started. It is a good idea to bring this extra information into the imputation model to stabilize the solution. More precisely, we may restrict any imputations of `gen`, `phb` and `tv`

to the lowest possible category for those boys younger than 8 years. This can be achieved by

```
> post <- mice(boys, m = 1, maxit = 0)$post
> post["gen"] <- "ifdo(age<8, levels(gen)[1])"
> post["phb"] <- "ifdo(age<8, levels(phb)[1])"
> post["tv"] <- "ifdo(age<8, 1)"
> free <- mice(boys, m = 1, seed = 85444)
> restricted <- mice(boys, m = 1, post = post, seed = 85444)
```

Figure 5.4 compares the scatterplot of genital development against age for the free and restricted solutions. Around infancy and early childhood, the imputation generated under the free solution are clearly unrealistic due to the severe extrapolation of the data between the ages 0–8 years. The restricted solution remedies this situation by requiring that pubertal development does not start before the age of 8 years.

The post-processing facility provides almost limitless possibilities to customize the imputed values. For example, we could reset the imputed value in some subset of the missing data to `NA`, thus imputing only some of the variables. Of course, appropriate care is needed when using this partially imputed variable later on as a predictor. Another possibility is to add or multiply the imputed data by a given constant in the context of a sensitivity analysis for nonignorable missing data mechanisms (see Section 3.9). More generally, we might re-impute some entries in the dataset depending on their current value, thus opening up possibilities to specify methods for nonignorable missing data.

5.4.5 Compositional data ♠

Sometimes we know that a set of variables should add up to a given total. If one of the additive terms is missing, we can directly calculate its value with certainty by deducting the known terms from the total. This is known as deductive imputation (De Waal et al., 2011). If two additive terms are missing, imputing one of these terms uses the available one degree of freedom, and hence implicitly determines the other term. Data of this type are known as compositional data, and they occur often in household and business surveys. Imputation of compositional data has only recently received attention (Tempelman, 2007; Hron et al., 2010; De Waal et al., 2011).

This section suggests a basic method for imputing compositional data. Let $Y_{123} = Y_1 + Y_2 + Y_3$ be the known total score of the three variables Y_1, Y_2 and Y_3. We assume that Y_3 is complete and that Y_1 and Y_2 are jointly missing or observed. The problem is to create multiple imputations in Y_1 and Y_2 such that the sum of Y_1, Y_2 and Y_3 equals a given total Y_{123}, and such that parameters estimated from the imputed data are unbiased and have appropriate coverage.

Since Y_3 is known, we write $Y_{12} = Y_{123} - Y_3$ for the sum score $Y_1 + Y_2$. The key to the solution is to find appropriate values for the ratio $P_1 = Y_1/Y_{12}$,

or equivalently for $(1 - P_1) = Y_2/Y_{12}$. Let $P(P_1|Y_1^{\text{obs}}, Y_2^{\text{obs}}, Y_3, X)$ denote the posterior distribution of P_1, which is possibly dependent on the observed information. For each incomplete record, we make a random draw $\dot{P}_1$ from this distribution, and calculate imputations for Y_1 as $\dot{Y}_1 = \dot{P}_1 Y_{12}$. Likewise, imputations for Y_2 are calculated by $\dot{Y}_2 = (1 - \dot{P}_1)Y_{12}$. It is easy to show that $\dot{Y}_1 + \dot{Y}_2 = Y_{12}$, and hence $\dot{Y}_1 + \dot{Y}_2 + Y_3 = Y_{123}$, as required.

The way in which the posterior should be specified is still an open problem. In this section we apply standard predictive mean matching. We study the properties of the method by a small simulation study. The first step is to create an artificial dataset with known properties as follows:

```
> set.seed(43112)
> n <- 400
> Y1 <- sample(1:10, size = n, replace = TRUE)
> Y2 <- sample(1:20, size = n, replace = TRUE)
> Y3 <- 10 + 2 * Y1 + 0.6 * Y2 + sample(-10:10,
      size = n, replace = TRUE)
> Y <- data.frame(Y1, Y2, Y3)
> Y[1:100, 1:2] <- NA
> md.pattern(Y)
    Y3  Y1  Y2
300  1   1   1   0
100  1   0   0   2
     0 100 100 200
```

Thus, `Y` is a 400×3 dataset with 300 complete records and with 100 records in which both Y_1 and Y_2 are missing. Next, define three auxiliary variables that are needed for imputation:

```
> Y123 <- Y1 + Y2 + Y3
> Y12 <- Y123 - Y[, 3]
> P1 <- Y[, 1]/Y12
> data <- data.frame(Y, Y123, Y12, P1)
```

where the naming of the variables corresponds to the total score Y_{123}, the sum score Y_{12} and the ratio P_1.

The imputation model specifies how Y_1 and Y_2 depend on P_1 and Y_{12} by means of passive imputation. The predictor matrix specifies that only Y_3 and Y_{12} may be predictors of P_1 in order to avoid linear dependencies.

```
> ini <- mice(data, maxit = 0, m = 10, print = FALSE,
      seed = 21772)
> meth <- ini$meth
> meth["Y1"] <- "~I(P1*Y12)"
> meth["Y2"] <- "~I((1-P1)*Y12)"
> meth["Y12"] <- "~I(Y123-Y3)"
> pred <- ini$pred
```

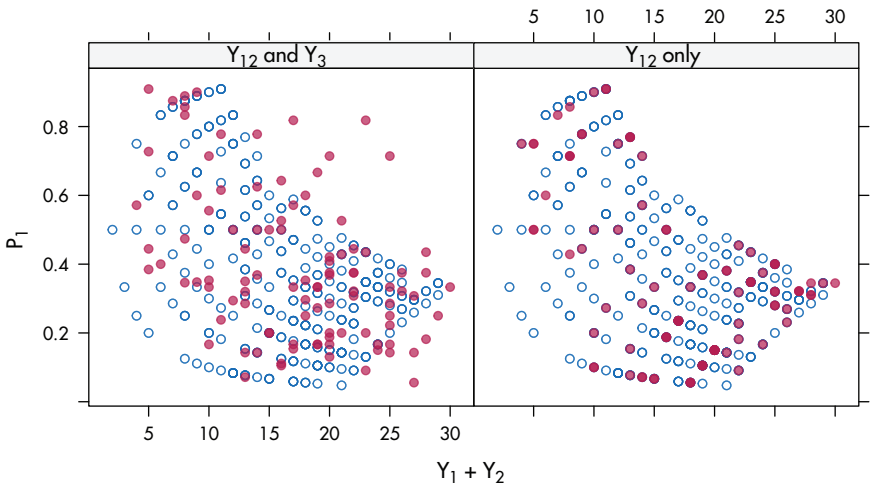


Figure 5.5: Distribution of P_1 (relative contribution of Y_1 to $Y_1 + Y_2$) in the observed and imputed data at different levels of Y_1+Y_2. The strong geometrical shape in the observed data is partially reproduced in the model that includes Y_3.

```
> pred["P1", ] <- 0
> pred[c("P1"), c("Y12", "Y3")] <- 1
> imp1 <- mice(data, meth = meth, pred = pred, m = 10,
      print = FALSE)
```

The code `I(P1*Y12)` calculates Y_1 as the product of P_1 and Y_{12}, and so on. The pooled estimates are calculated as

```
> round(summary(pool(with(imp1, lm(Y3 ~ Y1 + Y2))))[,
      1:2], 2)
             est   se
(Intercept) 9.58 1.00
Y1          2.00 0.11
Y2          0.61 0.05
```

The estimates are reasonably close to their true values of 10, 2 and 0.6, respectively. A small simulation study with these data using 100 simulations and $m = 10$ revealed average estimates of 9.94 (coverage 0.96), 1.95 (coverage 0.95) and 0.63 (coverage 0.91). Though not perfect, the estimates are close to the truth, while the data adhere to the summation rule.

Figure 5.5 shows where the solution might be further improved. The distribution of P_1 in the observed data is strongly patterned. This pattern is only partially reflected in the imputed $\dot{P}_1$ after predictive mean matching on both

Y_{12} and Y_3. It is possible to imitate the pattern perfectly by removing Y_3 as a predictor for P_1. However, this introduces bias in the parameter estimates. Evidently, some sort of compromise between these two options might further remove the remaining bias. This is an area for further research.

For a general missing data pattern, the procedure can be repeated for all pairs $(Y_j, Y_{j'})$ that have missing data. First create a consistent starting imputation that adheres to the rule of composition, then apply the above method to pairs $(Y_j, Y_{j'})$ that belong to the composition. This algorithm is a variation on the MICE algorithm with iterations occurring over pairs of variables rather than separate variables. The properties of this method have not yet been explored.

5.4.6 Quadratic relations ♠

One way to analyze nonlinear relations by a linear model is to include quadratic or cubic versions of the explanatory variables into the model. Creating imputed values under such models poses some challenges. Current imputation methodology either preserves the quadratic relation in the data and biases the estimates of interest, or provides unbiased estimates but does not preserve the quadratic relation (Von Hippel, 2009). White et al. (2011b) described three approaches:

1. *Linear passive*: Y is imputed under a linear model, and Y^2 is imputed passively by taking the square of Y.
2. *Improved passive*: The method is identical to "linear passive," with the linear imputation model replaced by predictive mean matching.
3. *Just another variable (JAV)*: The method imputes Y and Y^2 as separate variables, ignoring the deterministic relation (cf. Section 5.4.1).

White et al. (2011b) observed that "linear passive" biases the regression coefficient of Y^2 toward zero. The imputation model wrongly assumes a linear relation, and is thus uncongenial. The "improved passive" method is closer to being congenial, and is therefore to be preferred over "linear passive." JAV is congenial, but highly misspecified.

Summing up, it seems that we either have a congenial but misspecified model, or an uncongenial model that is specified correctly. The remainder of the section describes a new approach that aims to resolve this problem.

The model of scientific interest is

$$X = \alpha + Y\beta_1 + Y^2\beta_2 + \epsilon \tag{5.1}$$

with $\epsilon \sim N(0, \sigma^2)$. We assume that X is complete, and that $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ is partially missing. The problem is to find imputations for Y such that estimates of α, β_1, β_2 and σ^2 based on the imputed data are unbiased, while ensuring that the quadratic relation between Y and Y^2 will hold in the imputed data.

Define the *polynomial combination* Z as $Z = Y\beta_1 + Y^2\beta_2$ for some β_1 and β_2. The idea is to impute Z instead of Y and Y^2, followed by a decomposition of the imputed data Z into components Y and Y^2. Imputing Z reduces the multivariate imputation problem to a univariate problem, which is easier to manage. Under the assumption that $P(X, Z)$ is multivariate normal, we can impute the missing part of Z by Algorithm 3.1. In cases where a normal residual distribution is suspect, we replace the linear model by predictive mean matching. The next step is to decompose Z into Y and Y^2. Under model 5.1 the value Y has two roots:

$$Y_- = -\tfrac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} + \beta_1\right) \tag{5.2}$$

$$Y_+ = \tfrac{1}{2\beta_2}\left(\sqrt{4\beta_2 Z + \beta_1^2} - \beta_1\right) \tag{5.3}$$

where we assume that the discriminant $4\beta_2 Z + \beta_1^2$ is larger than zero. For a given Z, we can take either $Y = Y_-$ or $Y = Y_+$, and square it to obtain Y^2. Either root is consistent with $Z = Y\beta_1 + Y^2\beta_2$, but the choice among these two options requires care. Suppose we choose Y_- for all Z. Then all Y will correspond to points located on the left arm of the parabolic function. The minimum of the parabola is located at $Y_{\min} = -\beta_1/2\beta_2$, so all imputations will occur in the left-hand side of the parabola. This is probably not intended.

The choice between the roots is made by random sampling. Let V be a binary random variable defined as 1 if $Y > Y_{\min}$, and as 0 if $Y \leq Y_{\min}$. Let us model the probability $P(V = 1)$ by logistic regression as

$$\text{logit}(P(V = 1)) = X\psi_X + Z\psi_Z + XZ\psi_{YZ} \tag{5.4}$$

where the ψs are parameters in the logistic regression model. Under the assumption of ignorability, we calculate the predicted probability $P(V = 1)$ from X_{mis} and Z_{mis}. As a final step, a random draw from the binomial distribution is made, and the corresponding (negative or positive) root is selected as the imputation. This is repeated for each missing value.

Algorithm 5.1 provides a detailed overview of all steps involved. The imputations $\dot{Z}$ satisfy $\dot{Z} = \dot{Y}\hat{\beta}_1 + \dot{Y}^2\hat{\beta}_2$, as required. Work is under way to investigate the properties of the method. It might be of interest to generalize this method to polynomial bases of higher order.

5.5 Algorithmic options

5.5.1 Visit sequence

The default MICE algorithm imputes incomplete columns in the data from left to right. Theoretically, the visit sequence of the MICE algorithm is irrel-

Algorithm 5.1: Multiple imputation of quadratic terms.♠

1. Calculate Y^2_{obs} for the observed Y.
2. Multiply impute Y_{mis} and Y^2_{mis} as if they were unrelated by linear regression or predictive mean matching, resulting in imputations $\dot{Y}$ and $\dot{Y}^2$.
3. Estimate $\hat{\beta}_1$ and $\hat{\beta}_2$ by pooled linear regression of X given $Y = (Y_{\text{obs}}, \dot{Y})$ and $Y^2 = (Y^2_{\text{obs}}, \dot{Y}^2)$.
4. Calculate the polynomial combination $Z = Y\hat{\beta}_1 + Y^2\hat{\beta}_2$.
5. Multiply impute Z_{mis} by linear regression or predictive mean matching, resulting in imputed $\dot{Z}$.
6. Calculate roots $\dot{Y}_-$ and $\dot{Y}_+$ given $\hat{\beta}_1$, $\hat{\beta}_2$ and $\dot{Z}$ using Equations 5.2 and 5.3.
7. Calculate the value on the horizontal axis at the parabolic minimum/maximum $Y_{\text{min}} = -\hat{\beta}_1/2\hat{\beta}_2$.
8. Calculate $V_{\text{obs}} = 0$ if $Y_{\text{obs}} \leq Y_{\text{min}}$, else $V_{\text{obs}} = 1$.
9. Impute V_{mis} by logistic regression of V given X, Z and XZ, resulting in imputed $\dot{V}$.
10. If $\dot{V} < 0$ then assign $\dot{Y} = \dot{Y}_-$, else set $\dot{Y} = \dot{Y}_+$.
11. Calculate $\dot{Y}^2$.

evant as long as each column is visited often enough, though some schemes are more efficient than others. In practice, there are small order effects of the MICE algorithm, where the parameter estimates depend on the sequence of the variables. To date, there is little evidence that this matters in practice, even for clearly incompatible imputation models (Van Buuren et al., 2006). For monotone missing data, convergence is immediate if variables are ordered according to their missing data rate. Rather than reordering the data, it is more convenient to change the visit sequence of the algorithm by the `visitSequence` argument. In its basic form, the `visitSequence` argument is a vector of integers in the range `1:ncol(data)` of arbitrary length, specifying the sequence of column numbers for one iteration of the algorithm. Any given column may be visited more than once within the same iteration, which can be useful to ensure proper synchronization among variables.

Consider the `mids` object `imp.int` created in Section 5.4.3. The visit sequence is

```
> imp.int$vis
   hgt    wgt     hc    gen    phb     tv    reg wgt.hc
     2      3      5      6      7      8      9     10
```

If the `visitSequence` is not specified, the `mice()` function imputes the data from left to right. Thus here `wgt.hc` is calculated after `reg` is imputed, so at this point `wgt.hc` is synchronized with both `wgt` and `hc`. Note, however, that `wgt.hc` is not synchronized with `wgt` and `hc` when imputing `pub`, `gen`, `tv` or `reg`, so `wgt.hc` is not representing the current interaction effect. This could result in wrong imputations. We can correct this by including an extra visit to `wgt.hc` after `wgt` or `hc` has been imputed as follows:

```
> vis <- c(2, 3, 5, 10, 6:9)
> expr <- expression((wgt - 40) * (hc - 50))
> boys$wgt.hc <- with(boys, eval(expr))
> imp.int2 <- mice(boys, m = 1, max = 1, vis = vis,
      meth = imp.int$meth, pred = imp.int$pred,
      seed = 23390)
 iter imp variable
  1   1  hgt  wgt  hc  wgt.hc  gen  phb  tv  reg
```

When the missing data pattern is close to monotone, convergence may be speeded by visiting the columns in increasing order of the number of missing data. We can specify this order by the `"monotone"` keyword as

```
> imp.int2 <- mice(boys, m = 1, max = 1, vis = "monotone",
      meth = imp.int$meth, pred = imp.int$pred,
      seed = 23390)
 iter imp variable
  1   1  reg  wgt  hgt  bmi  hc  wgt.hc  gen  phb  tv
```

5.5.2 Convergence

There is no clear-cut method for determining when the MICE algorithm has converged. It is useful to plot one or more parameters against the iteration number. The mean and variance of the imputations for each parallel stream can be plotted by

```
> imp <- mice(nhanes, seed = 62006, maxit = 20,
      print = FALSE)
> plot(imp)
```

which produces Figure 5.6. On convergence, the different streams should be freely intermingled with one another, without showing any definite trends. Convergence is diagnosed when the variance between different sequences is no larger than the variance within each individual sequence.

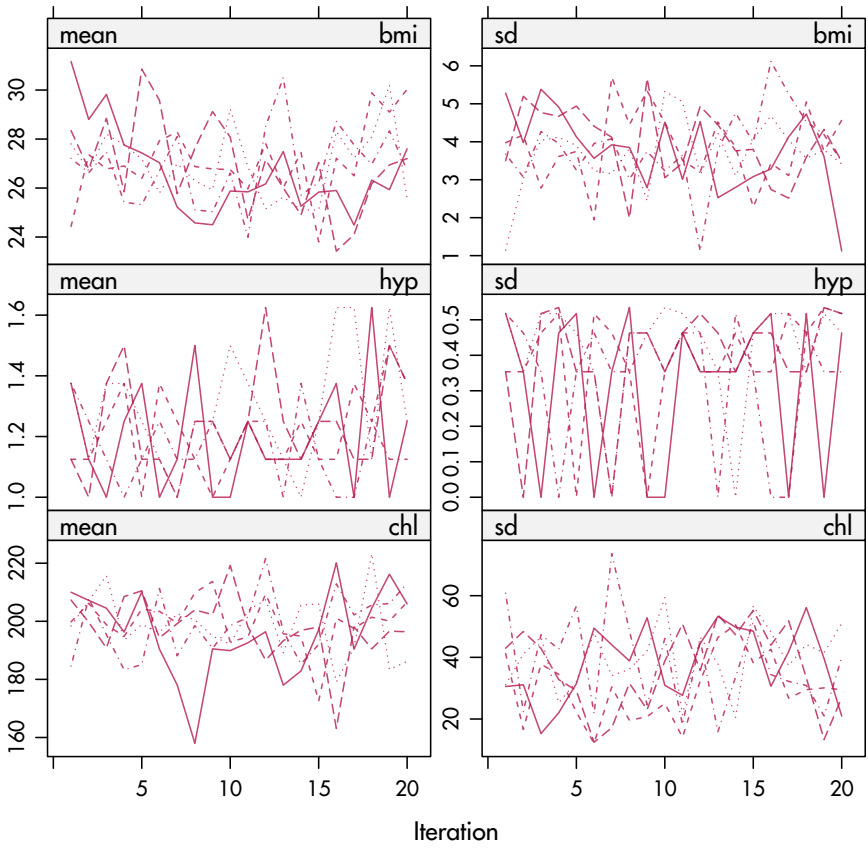


Figure 5.6: Mean and standard deviation of the synthetic values plotted against iteration number for the imputed `nhanes` data.

Inspection of the streams may reveal particular problems of the imputation model. A pathological case of non-convergence occurs with the following code:

```
> ini <- mice(boys, max = 0, print = FALSE)
> meth <- ini$meth
> meth["bmi"] <- "~I(wgt/(hgt/100)^2)"
> imp.bmi1 <- mice(boys, meth = meth, maxit = 20,
      seed = 60109)
```

Convergence is problematic because imputations of `bmi` feed back into `hgt` and `wgt`. Figure 5.7 shows that the streams hardly mix and slowly resolve into a steady state. The problem is solved by breaking the feedback loop as follows:

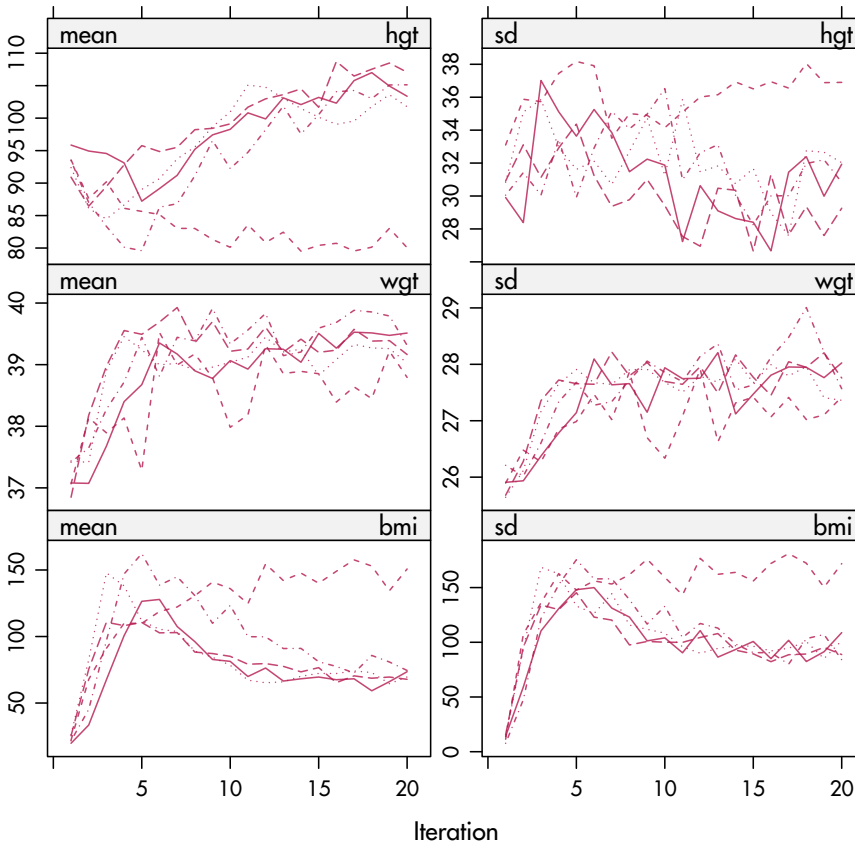


Figure 5.7: Non-convergence of the MICE algorithm caused by feedback of `bmi` into `hgt` and `wgt`.

```
> pred <- ini$pred
> pred[c("hgt", "wgt"), "bmi"] <- 0
> imp.bmi2 <- mice(boys, meth = meth, pred = pred,
      maxit = 20, seed = 60109)
```

Figure 5.8 is the resulting plot for the same three variables. There is little trend and the streams mingle well.

The default `plot()` function for `mids` objects plots the mean and variance of the imputations. While these parameters are informative for the behavior of the MICE algorithm, they may not always be the parameter of greatest interest. It is easy to replace the mean and variance by other parameters, and monitor these. Schafer (1997, p. 129–131) suggested monitoring the "worst linear function" of the model parameters, i.e., a combination of parameters

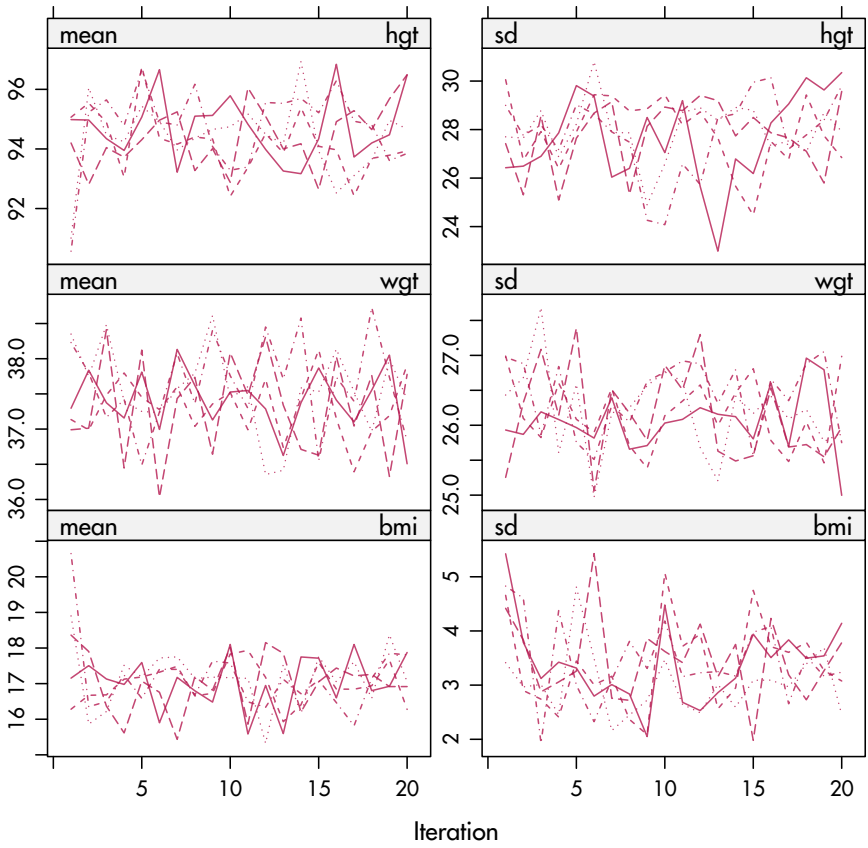


Figure 5.8: Healthy convergence of the MICE algorithm for `hgt`, `wgt` and `bmi`.

that will experience the most problematic convergence. If convergence can be established for this parameter, then it is likely that convergence will also be achieved for parameters that converge faster. Alternatively, we may monitor some statistic of scientific interest, e.g., a correlation or a proportion. See Section 7.4.3 for an example.

It is possible to use formal convergence statistics. Several expository reviews are available that assess convergence diagnostics for MCMC methods (Cowles and Carlin, 1996; Brooks and Gelman, 1998; El Adlouni et al., 2006). Cowles and Carlin (1996) conclude that "automated convergence monitoring (as by a machine) is unsafe and should be avoided." No method works best in all circumstances. The consensus is to assess convergence with a combination

of tools. The added value of using a combination of convergence diagnostics for missing data imputation has not yet been systematically studied.

5.6 Diagnostics

Diagnostics for statistical models are procedures to find departures from the assumptions underlying the model. Model evaluation is a huge field in which many special tools are available, e.g., Q-Q plots, residual and influence statistics, formal statistical tests, information criteria and posterior predictive checks. In principle, all these techniques can be applied to evaluate the imputation model. Conventional model evaluation concentrates on the fit between the data and the model. In imputation it is often more informative to focus on *distributional discrepancy*, the difference between the observed and imputed data. The next section illustrates this with an example.

5.6.1 Model fit versus distributional discrepancy

The MICE algorithm fits the imputation model to the records with observed Y_j (Y_j^{obs}), and applies the fitted model to generate imputations for the records with unobserved Y_j (Y_j^{mis}). The fit of the imputation model to the data can thus be studied from Y_j^{obs}.

The worm plot is a diagnostic tool to assess the fit of a nonlinear regression (Van Buuren and Fredriks, 2001; Van Buuren, 2007b). In technical terms, the worm plot is a detrended Q-Q plot conditional on a covariate. The model fits the data if the worms are close to the horizontal axis.

Figure 5.9 is the worm plot calculated from imputed data after predictive mean matching. The fit between the observed data and the imputation model is bad. The blue points are far from the horizontal axis, especially for the youngest children. The shapes indicate that the model variance is much larger than the data variance. In contrast to this, the red and blue worms are generally close, indicating that the distributions of the imputed and observed body weights are similar. Thus, despite the fact that the model does not fit the data, the distributions of the observed and imputed data are similar. This distributional similarity is more relevant for the final inferences than model fit per se.

5.6.2 Diagnostic graphs

One of the best tools to assess the plausibility of imputations is to study the discrepancy between the observed and imputed data. The idea is that good imputations have a distribution similar to the observed data. In other words,

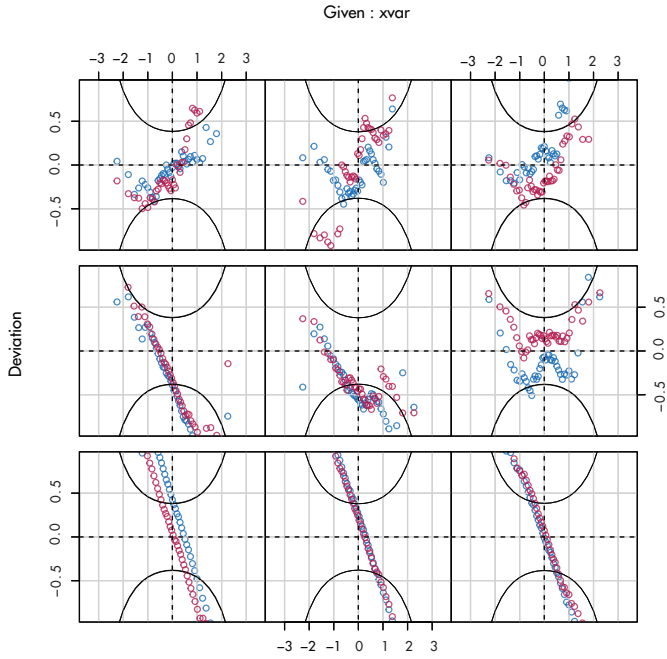


Figure 5.9: Worm plot of the predictive mean matching imputations for body weight. Different panels correspond to different age ranges. Though the imputation model has a bad fit in many age groups, the distributions of the observed and imputed data often match up very well.

the imputations could have been real values had they been observed. Except under MCAR, the distributions do not need to be identical, since strong MAR mechanisms may induce systematic differences between the two distributions. However, any dramatic differences between the imputed and observed data should certainly alert us to the possibility that something is wrong.

This book contains many colored figures that emphasize the relevant contrasts. Graphs allow for a quick identification of discrepancies of various types:

- the points have different means (Figure 2.2);
- the points have different spreads (Figures 1.1, 1.2 and 1.4);
- the points have different scales (Figure 4.4);
- the points have different relations (Figure 5.1);
- the points do not overlap and they defy common sense (Figure 5.4).

Differences between the densities of the observed and imputed data may suggest a problem that needs to be further checked. The `mice` package (V2.6 and

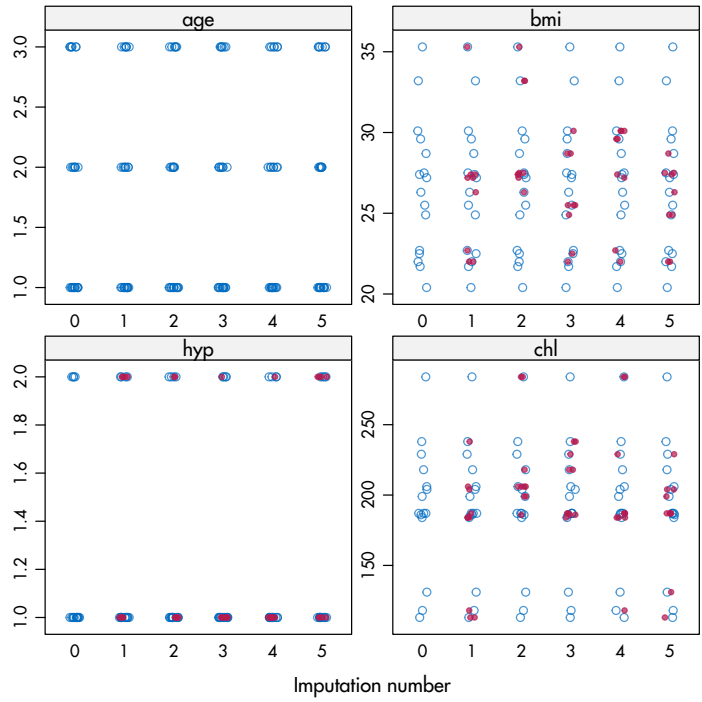


Figure 5.10: A stripplot of the multiply imputed `nhanes` data with $m = 5$.

up) contains several graphic functions that can be used to gain insight into the correspondence of the observed and imputed data: `bwplot()`, `stripplot()`, `densityplot()` and `xyplot()`.

The `stripplot()` function produces the individual points for numerical variables per imputation as in Figure 5.10 by

```
> imp <- mice(nhanes, seed = 29981)
> stripplot(imp, pch = c(1, 20))
```

The stripplot is useful to study the distributions in datasets with a low number of data points. For large datasets it is more appropriate to use the function `bwplot()` that produces side-by-side box-and-whisker plots for the observed and synthetic data.

The `densityplot()` function produces Figure 5.11 by

```
> densityplot(imp)
```

The figure shows kernel density estimates of the imputed and observed data. In this case, the distributions match up well.

Interpretation is more difficult if there are discrepancies. Such discrepancies

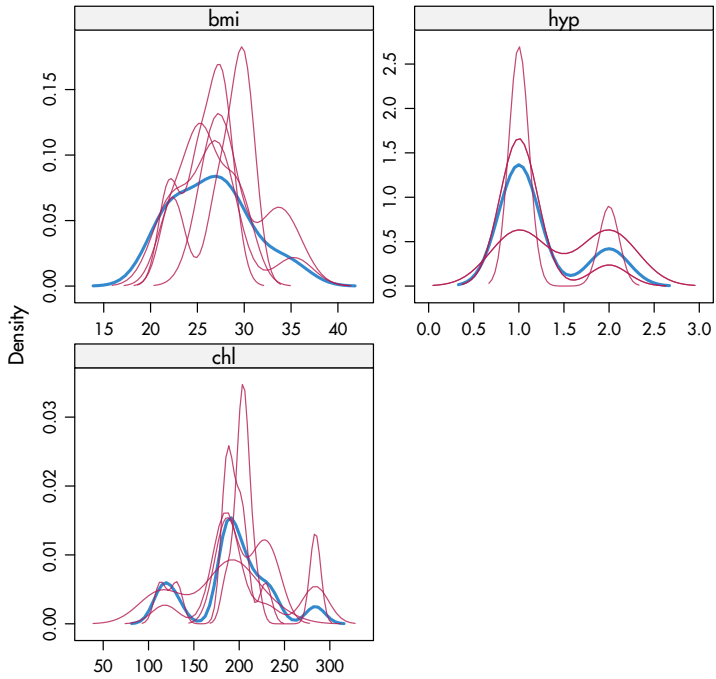


Figure 5.11: Kernel density estimates for the marginal distributions of the observed data (blue) and the $m = 5$ densities per variable calculated from the imputed data (thin red lines).

may be caused by a bad imputation model, by a missing data mechanism that is not MCAR or by a combination of both. Raghunathan and Bondarenko (2007) proposed a more refined diagnostic tool that aims to compare the distributions of observed and imputed data conditional on the missingness probability. The idea is that under MAR the conditional distributions should be similar if the assumed model for creating multiple imputations has a good fit. An example is created as follows:

```
> fit <- with(imp, glm(ici(imp) ~ age + bmi + hyp +
      chl, family = binomial))
> ps <- rep(rowMeans(sapply(fit$analyses, fitted.values)),
      imp$m)
> xyplot(imp, bmi ~ ps | .imp, xlab = "Response probability",
      ylab = "BMI", pch = c(1, 19))
```

These statements first model the probability of each record being incomplete as a function of all variables in each imputed dataset. The probabilities (propensities) are then averaged over the imputed datasets to obtain stability. Fig-

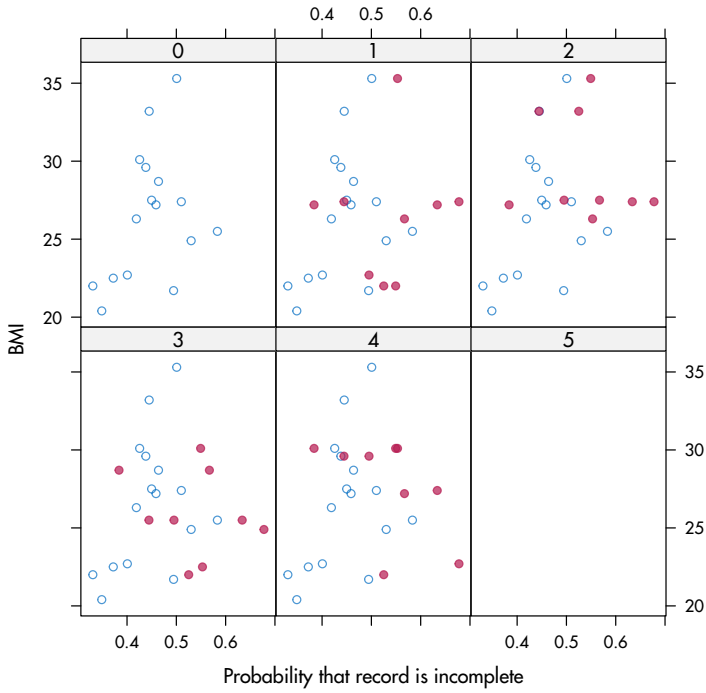


Figure 5.12: BMI against missingness probability for observed and imputed values.

ure 5.12 plots BMI against the propensity score in each dataset. Observe that the imputed data points are somewhat shifted to the right. In this case, the distributions of the blue and red points are quite similar, as expected under MAR.

Realize that the comparison is only as good as the propensity score. If important predictors are omitted from the response model, then we may not be able to see the potential misfit. In addition, it could be useful to investigate the residuals of the regression of BMI on the propensity score. See Van Buuren and Groothuis-Oudshoorn (2011) on a technique for how to calculate and plot the relevant quantities.

Compared to conventional diagnostic methods, imputation comes with the advantage that we can directly compare the observed and imputed values. The marginal distributions of the observed and imputed data may differ because the missing data are MAR or MNAR. The diagnostics tell us in what way they differ, and hopefully also suggest whether these differences are expected and sensible in light of what we know about the data. Under MAR, any distributions that are conditional on the missing data process should be the same. If our diagnostics suggest otherwise (e.g., the blue and red points are

very different), there might be something wrong with the imputations that we created. Alternatively, it could be the case that the observed differences are justified, and that the missing data process is MNAR. The art of imputation is to distinguish between these two explanations.

5.7 Conclusion

Multiple imputation is not a quick automatic fix. Creating good imputations requires substantive knowledge about the data paired with some healthy statistical judgement. Impute close to the data. Real data are richer and more complex than the statistical models applied to them. Ideally, the imputed values should look like real data in every respect, especially if multiple models are to be fit on the imputed data. Keep the following points in mind:

- Plan time to create the imputed datasets. As a rule of thumb, reserve for imputation 5% of the time needed to collect the data;
- Check the modeling choices in Section 5.1. Though the software defaults are often reasonable, they may not work for the particular data;
- Use MAR as a starting point using the strategy outlined in Section 5.2;
- Choose the imputation methods and set the predictors using the strategies outlined in Section 5.3;
- If the data contain derived variables that are not needed for imputation, impute the originals and calculate the derived variables afterward;
- Use passive imputation if you need the derived variables during imputation. Carefully specify the predictor matrix to avoid feedback loops. See Section 5.4;
- Monitor convergence of the MICE algorithm for aberrant patterns, especially if the rate of missing data is high or if there are dependencies in the data. See Sections 4.5.5 and 5.5.2;
- Make liberal use of diagnostic graphs to compare the observed and the imputed data. Convince yourself that the imputed values could have been real data, had they not been missing. See Section 5.6.

5.8 Exercises

1. *Worm plot for normal model.* Repeat the imputations in Section 5.6.1 using the linear normal model for the numerical variables. Draw the worm plot.

   - Does the imputation model for `wgt` fit the observed data? If not, describe in which aspects they differ.
   - Does the imputation model for `wgt` fit the imputed data? If not, describe in which aspects they differ.
   - Are there striking differences between your worm plot and Figure 5.9? If so, describe.
   - Which imputation model do you prefer? Why?

2. *Defaults.* Select a real dataset that is familiar to you and that contains at least 20% missing data. Impute the data with `mice()` under all the default settings.

   - Inspect the streams of the MICE algorithm. Does the sampler appear to converge?
   - Extend the analysis with 20 extra iterations using `mice.mids()`. Does this affect your conclusion about convergence?
   - Inspect the data with diagnostic plots for univariate data. Are the univariate distributions of the observed and imputed data similar? Do you have an explanation why they do (or do not) differ?
   - Inspect the data with diagnostic plots for the most interesting bivariate relations. Are the relations similar in the observed and imputed data? Do you have an explanation why they do (or do not) differ?
   - Consider each of the seven default choices in turn. Do you think the default is appropriate for your data? Explain why.
   - Do you have particular suggestions for improved choices? Which?
   - Implement one of your suggestions. Do the results now look more plausible or realistic? Explain.

Chapter 6

Analysis of imputed data

Creating plausible imputations is the most challenging activity in multiple imputation. Though the steps needed to derive the final statistical inferences are relatively straightforward, they are not entirely without problems. This chapter describes techniques to pool non-normal quantities. In addition, the chapter reviews methods to deal with variable selection and model optimism on multiply imputed data.

6.1 What to do with the imputed data?

6.1.1 Averaging and stacking the data

Novices at multiple imputation are sometimes tempted to average the multiply imputed data, and analyze the averaged data as if it were complete. This method yields incorrect standard errors, confidence intervals, and p-values, and thus should not be used. The reason is that the procedure ignores the between-imputation variability, and hence the procedure shares all the drawbacks of single imputation. See Section 1.3.

A variation on this theme is to stack the data, thus creating $m \times n$ complete records. Each record is weighted by a factor $1/m$, so that the total sample size is equal to n. The statistical analysis amounts to performing a, for example, weighted linear regression. Though this procedure yields unbiased point estimates, the resulting inferences will be incorrect since the implied sample size is too large if there are missing data.

We could consider alternatives to decrease the weights proportional to the amount of missing data. Wood et al. (2008) suggested some possibilities in a different context; however, the properties of the final inferences are unknown. Kim and Fuller (2004) proposed a method called fractional imputation, in which the imputed values are weighted explicitly by design-based weights. While this method yields appropriate variance estimates in particular cases, it is unclear how general this method is, or whether new complexities will arise if the weights themselves are estimates.

If the scientific interest is solely restricted to the point estimate, then the stacked imputed data can be validly used to obtain a quick unbiased estimate

for linear models. Be aware that routine methods for calculating test statistics, confidence intervals, or p-values will provide invalid answers if applied to the stacked imputed data.

Creating and analyzing a stacked imputed dataset is easy to do:

```
> imp <- mice(nhanes, print = FALSE, seed = 55152)
> stacked <- complete(imp, "long")
> fit <- lm(chl ~ bmi + age, data = stacked)
> coef(fit)
(Intercept)         bmi         age
      -6.15        5.41       31.70
```

Equal weights are needed to obtain appropriate point estimates. While the estimated regression coefficients are unbiased, we cannot trust the standard errors, t-values, and so on.

6.1.2 Repeated analyses

The appropriate way to analyze multiply imputed data is to perform the complete data analysis on each imputed dataset separately. In `mice` we can use the `with()` command for this purpose. This function takes two main arguments. The first argument of the call is a `mids` object produced by `mice()`. The second argument is an expression that is to be applied to each completed dataset. The `with()` function implements the following loop ($\ell = 1, \ldots, m$):

1. it creates the ℓ-th imputed dataset
2. it runs the expression on the imputed dataset
3. it stores the result in the list `fit$analyses`

For example, we fit a regression model to each dataset and print out the estimates from the first and second completed datasets by

```
> fit <- with(imp, lm(chl ~ bmi + age))
> coef(fit$analyses[[1]])
(Intercept)         bmi         age
     -29.18        5.63       43.20
> coef(fit$analyses[[2]])
(Intercept)         bmi         age
     -36.57        6.77       25.63
```

Note that the estimates differ from each other because of the uncertainty created by the missing data. Applying the standard pooling rules is done by

```
> est <- pool(fit)
> summary(est)
```

```
               est    se     t   df Pr(>|t|)    lo 95 hi 95
(Intercept) -10.99 68.71 -0.16 6.62   0.8777 -175.359 153.4
bmi           5.55  2.24  2.47 6.87   0.0433    0.223  10.9
age          32.11 11.91  2.70 5.57   0.0385    2.415  61.8
            nmis   fmi lambda
(Intercept)   NA 0.585  0.476
bmi            9 0.572  0.463
age            0 0.646  0.539
```

which shows that the estimates are fairly close to the estimates calculated in the previous section.

Any R expression produced by `expression()` can be evaluated on the multiply imputed data. For example, suppose we want to calculate the difference in frequencies between categories 1 and 2 of `hyp`. This is conveniently done by the following statements:

```
> expr <- expression(freq <- table(hyp), freq[1] -
      freq[2])
> fit <- with(imp, eval(expr))
> unlist(fit$analyses)
 1  1  1  1  1
13 15 15 15 13
```

All the major software packages nowadays have ways to execute the m repeated analyses to the imputed data.

6.2 Parameter pooling

6.2.1 Scalar inference of normal quantities

Section 2.4 describes Rubin's rules for pooling the results from the m complete data analyses. These rules are based on the assumption that the parameter estimates $\hat{Q}$ are normally distributed around the population value Q with a variance of U. Many types of estimates are approximately normally distributed, e.g., means, standard deviations, regression coefficients, proportions and linear predictors. Rubin's pooling rules can be applied directly to such quantities (Schafer, 1997; Marshall et al., 2009).

6.2.2 Scalar inference of non-normal quantities

How should we combine quantities with non-normal distributions: correlation coefficients, odds ratios, relative risks, hazard ratios, measures of explained variance and so on? The quality of the pooled estimate and the confidence intervals can be improved when pooling is done in a scale for which the

Table 6.1: Suggested transformations toward normality for various types of statistics. The transformed quantities can be pooled by Rubin's rules.

Statistic	Transformation	Source
Correlation	Fisher z	Schafer (1997)
Odds ratio	Logarithm	Agresti (1990)
Relative risk	Logarithm	Agresti (1990)
Hazard ratio	Logarithm	Marshall et al. (2009)
Explained variance R^2	Fisher z on root	Harel (2009)
Survival probabilities	Complementary log-log	Marshall et al. (2009)
Survival distribution	Logarithm	Marshall et al. (2009)

distribution is close to normal. Thus, transformation toward normality and back-transformation into the original scale improves statistical inference.

As an example, consider transforming a correlation coefficient ρ_ℓ for $\ell = 1, \ldots, m$ toward normality using the Fisher z transformation

$$z_\ell = \frac{1}{2} \ln \frac{1 + \rho_\ell}{1 - \rho_\ell} \tag{6.1}$$

For large samples, the distribution of z_ℓ is normal with variance $\sigma^2 = 1/(n-3)$. It is straightforward to calculate the pooled correlation $\bar{z}$ and its variance by Rubin's rules. The result can be back-transformed by the inverse Fisher transformation

$$\bar{\rho} = \frac{e^{2\bar{z}} - 1}{e^{2\bar{z}} + 1} \tag{6.2}$$

The confidence interval of $\bar{\rho}$ is calculated in the z-scale as usual, and then back-transformed by Equation 6.2.

Table 6.1 suggests transformations toward approximate normality for various types of statistics. There are quantities for which the distribution is complex or unknown. Examples include the Cramér C statistic (Brand, 1999) and the discrimination index (Marshall et al., 2009). Ideally, the entire sampling distribution should be pooled in such cases, but the corresponding pooling methods have yet to be developed. The current advice is to search for ad hoc transformations to make the sampling distribution close to normality, and then apply Rubin's rules.

6.3 Statistical tests for multiple imputation

Special pooling procedures have been developed for several statistical tests for multiply imputed data: the Wald test, the likelihood ratio test, and the χ^2-test. The next sections describe these tests. These sections are more technical than the rest of the book, and are included primarily for completeness.

6.3.1 Wald test ♠

The multivariate Wald test requires an estimate of the total $k \times k$ variance-covariance matrix T. For small m the estimate of the between-imputation variance B is unstable, and if $m \leq k$, it is not even full rank. Thus T could also be unreliable, especially if B makes up a substantial part of T.

Li et al. (1991b) proposed an estimate of T in which B and $\bar{U}$ are proportional to each other. A more stable estimate of the total variance is $\tilde{T} = (1 + \bar{r})\bar{U}$, which bypasses the need for B. The proportionality assumption is equivalent to assuming equal fractions of missing information, so $\bar{r}$ as defined by Equation 2.29 is considered to be a good overall measure. The test statistic is

$$D_w = (\bar{Q} - Q_0)'\tilde{T}^{-1}(\bar{Q} - Q_0)/k \tag{6.3}$$

where the p-value for D_w is

$$P_w = \Pr[F_{k,\nu_w} > D_w] \tag{6.4}$$

where F_{k,ν_w} is the F distribution with k and ν_w degrees of freedom, with

$$\nu_w = \begin{cases} 4 + (t-4)[1 + (1 - 2t^{-1})\bar{r}^{-1}]^2 & \text{if} \quad t = k(m-1) > 4 \\ t(1 + k^{-1})(1 + \bar{r}^{-1})^2/2 & \text{otherwise} \end{cases} \tag{6.5}$$

Although the assumption of equal fractions of missing information may seem limiting, Li et al. (1991b) provide very encouraging simulation results for situations where this assumption is violated. Except for some extreme cases, the level of the procedure was close to the nominal level. The loss of power from such violations was modest, approximately 10%.

The work of Li et al. (1991b) is based on large samples. Reiter (2007) developed a small-sample version for the degrees of freedom using ideas similar to Barnard and Rubin (1999). Reiter's ν_f spans several lines of text, and is not given here. A small simulation study conducted by Reiter showed marked improvement over the earlier formulation, especially in a smaller sample. It would be useful to see work that could confirm these results in more general settings.

In cases of practical interest, the test with $m > 3$ is insensitive to the assumption of equal fractions of missing information. Moreover, it is well calibrated, and suffers only modest loss of power. The test is recommended for application with low m, unless the fractions of information are large and variable. Calculation of the test requires access to $\bar{Q}$ and the variance-covariance matrices B and $\bar{U}$. These will typically be available in multiple imputation software.

6.3.2 Likelihood ratio test ♠

In some cases one cannot obtain the covariance matrices of the complete data estimates. This could be the case if the dimensionality of Q is high,

which can occur with partially classified contingency tables. The likelihood ratio test (Meng and Rubin, 1992) is designed to handle such situations. For large n the procedure is equivalent to the method of Section 6.3.1, and only requires calculation of the complete data log likelihood ratio statistic under two sets of parameter estimates in each multiply imputed dataset.

Let the vector Q contain the parameters of interest. We wish to test the hypothesis $Q = Q_0$ for some given Q_0. The usual scenario is that we have fit two models, one where Q can vary freely and one more restrictive model that constrains $Q = Q_0$. Let the corresponding estimates be denoted by $\bar{Q}$ and $\bar{Q}_0$.

In the ℓ-th imputed dataset Y_ℓ, we calculate the value of the log-likelihood functions $l(\bar{Q}|Y_\ell)$ and $l(\bar{Q}_0|Y_\ell)$ with $\ell = 1, \ldots, m$. The deviances between the models are defined as

$$\bar{d}_\ell = 2(l(\bar{Q}|Y_\ell) - l(\bar{Q}_0|Y_\ell)) \quad \text{for} \quad \ell = 1, \ldots, m \tag{6.6}$$

Likewise, for $\hat{Q}_\ell$ and $\hat{Q}_{0,\ell}$, the estimates that are optimal with respect to the ℓ-th imputed data, we calculate the deviance

$$d_\ell = 2(l(\hat{Q}_\ell|Y_\ell) - l(\hat{Q}_{0,\ell}|Y_\ell)) \quad \text{for} \quad \ell = 1, \ldots, m \tag{6.7}$$

The average deviances over the imputations are equal to $\bar{d} = 1/m \sum_{\ell=1}^{m} \bar{d}_\ell$ and $d = 1/m \sum_{\ell=1}^{m} d_\ell$, respectively. The test statistic proposed by Meng and Rubin (1992) is

$$D_l = \frac{\bar{d}}{k(1 + r_l)} \tag{6.8}$$

where

$$\bar{r}_l = \frac{m+1}{k(m-1)}(d - \bar{d}) \tag{6.9}$$

estimates the average relative increase in variance due to nonresponse. The estimate $\bar{r}_l$ is asymptotically equivalent to $\bar{r}_w$ from Equation 2.29. The p-value for D_l is equal to

$$P_l = \Pr[F_{k,\nu_l} > D_l] \tag{6.10}$$

where $\nu_l = \nu_w$. Alternatively, one may use Reiter's ν_f for small samples, using $\bar{r}_l$ as the estimate of the relative increase of the variance.

Schafer (1997, p. 118) argued that the best results will be obtained if the distribution of $\bar{Q}$ is approximately normal. One may transform the parameters to achieve normality, provided that appropriate care is taken to infer that the result is still within the allowable parameter space.

In complete data, the likelihood ratio test is often considered a better test than the Wald. This superiority does not hold however for the likelihood ratio test of this section. The likelihood ratio test as described here is asymptotically equivalent to the Wald test, and thus both tests share the same favorable properties. The choice between them is mostly a matter of convenience. If $\bar{U}$ and B are available, then the Wald test is usually preferred as there is no need for the likelihood function. The D_l statistic can be calculated without the need

for $\bar{U}$ and B, but requires evaluation of the likelihood function. The likelihood ratio test is a viable option if k is large (since $\bar{U}$ needs not be inverted), or in situations in which it is easier to calculate the likelihood than to obtain $\bar{U}$ and B.

6.3.3 χ^2-test ♠

Rubin (1987a, p. 87) and Li et al. (1991a) describe a procedure for pooling χ^2-statistics and its associated p-values. Suppose that χ^2_ℓ are test statistics obtained from the imputed data Y_ℓ, $\ell = 1, \ldots, m$. Let

$$\bar{\chi}^2 = \frac{1}{m} \sum_{\ell=1}^{m} \chi^2_\ell \tag{6.11}$$

be the average χ^2-statistic. The test statistic is

$$D_x = \frac{\frac{\bar{\chi}^2}{k} - \frac{m+1}{m-1}\bar{r}_x}{1 + \bar{r}_x} \tag{6.12}$$

where the relative increase of the variance is calculated as

$$\bar{r}_x = \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{\ell=1}^{m} \left(\sqrt{\chi^2_\ell} - \sqrt{\bar{\chi}^2}\right)^2 \tag{6.13}$$

The p-value corresponding to D_x is obtained as

$$P_x = \Pr[F_{k,\nu_x} > D_x] \tag{6.14}$$

where

$$\nu_x = k^{-3/m}(m-1)\left(1 + \frac{1}{\bar{r}_x^2}\right) \tag{6.15}$$

or Reiter's ν_f for small samples, using $\bar{r}_x$ as the estimate of the relative increase in variation due to the missing data.

The pooled χ^2-test can be used when k is large, if $\bar{U}$ and B cannot be retrieved, or if only χ^2-statistics are available. Compared to the other three methods, however, the results from the χ^2-test are considerably less reliable. The results were optimized for $m = 3$ and, unlike the other tests, do not necessarily improve for larger m. According to Li et al. (1991a) the true result could be within a range of one half to twice the obtained p-value. This test should only be used as a rough guide.

6.3.4 Custom hypothesis tests of model parameters ♠

Statistical intervals and tests based on scalar Q are used most widely, and are typically standard in multiple imputation software. Custom hypotheses

may be formulated and tested using linear contrasts of the model parameter estimates $\bar{Q}$.

Let C be an $h \times k$ matrix that specifies h linear combinations of the parameters in its rows, and let c be a specified column vector with h constants. The null hypotheses $C\bar{Q} = c$ can be tested on the multiply imputed data in the usual way. We calculate the custom estimate as $\bar{Q}^{(c)} = C\bar{Q}$ and its variance $T^{(c)} = \text{diag}(C\bar{Q}C')$. We find the p-value of the test as the probability

$$P_s = \Pr\left[F_{1,\nu} > \frac{(c - \bar{Q}^{(c)})^2}{T^{(c)}}\right] \tag{6.16}$$

where $F_{1,\nu}$ is an F-distribution with 1 and ν degrees of freedom. There seems to be no literature on how to calculate the degrees of freedom. An obvious choice would be to follow the method for the multivariate Wald statistic. It is not yet clear how well this method performs. Also, we could potentially improve by selecting only those parts of $\bar{U}$, B and T that contribute to the linear combination before calculating $\bar{r}$.

6.3.5 Computation

The `mice` package contains several functions for pooling parameter estimates. The standard function `pool()` extracts the fitted parameters from each repeated model, and applies the repeated scalar procedure (Section 2.4.2) to each parameter.

The multivariate Wald test and the likelihood ratio test are implemented in the function `pool.compare()`. This function takes two `mira` objects, where one model is nested within the other. We can apply the pooled Wald test to evaluate the difference between the current model and the intercept-only model by

```
> imp <- mice(nhanes2, seed = 23210, print = FALSE)
> fit <- with(imp, lm(bmi ~ age + chl))
> fit.restrict <- with(imp, lm(bmi ~ 1))
> res <- pool.compare(fit, fit.restrict)
> res$pvalue
       [,1]
[1,] 0.0182
```

Alternatively, the function can be used to apply the pooled likelihood ratio test, though currently only the likelihood function of the logistic regression model is implemented.

Confidence intervals and tests other than those preprogrammed can be calculated from the components of objects of class `mira`. To illustrate, we impute `nhanes2` and fit the linear model to predict `bmi` from the other variables. This is done as follows:

```
> est <- pool(fit)
> summary(est)
                 est     se     t    df Pr(>|t|)    lo 95
(Intercept) 17.7860 3.7542  4.74 10.68 0.000663   9.4933
age2        -4.3682 1.9134 -2.28 12.46 0.040709  -8.5202
age3        -5.7044 2.2716 -2.51  9.66 0.031600 -10.7900
chl          0.0593 0.0216  2.75 10.11 0.020223   0.0114
              hi 95 nmis   fmi lambda
(Intercept) 26.079   NA 0.392  0.288
age2        -0.216   NA 0.328  0.228
age3        -0.619   NA 0.432  0.326
chl          0.107   10 0.414  0.309
```

Note that `age` is coded in three groups, so there are two independent estimates, both of which are relative to the first level. We illustrate the use of three custom hypothesis tests on age. The first test assesses whether the difference between the estimates for categories `age2` and `age3` is significant. The second test assesses whether the difference between categories `age2` and `age3` is the same as the difference between categories `age1` and `age2`, which is the test for linearity. The third one tests the null hypothesis that the regression weight of `chl` is equal to 0.1.

```
> C <- matrix(c(c(0, 1, -1, 0), c(0, 2, -1, 0),
      c(0, 0, 0, 1)), nrow = 3, ncol = 4, byrow = TRUE)
> c <- c(0, 0, 0.1)
> q <- C %*% est$qbar
> u <- diag(C %*% est$ubar %*% t(C))
> t <- diag(C %*% est$t %*% t(C))
> C1 <- C
> C1[C1 != 0] <- 1
> df <- C1 %*% est$df
> d <- (q - c)^2/t
> pvalue <- 1 - pf(d, 1, df)
> pvalue
       [,1]
[1,] 0.6041
[2,] 0.4430
[3,] 0.0884
```

The interpretation is as follows. The difference between `age2` and `age3` is not significant, there is insufficient evidence to reject the linearity hypothesis for `age` and the average level of `chl` is not significantly different from 0.1. Though there are still some issues for determining the appropriate degrees of freedom, these types of custom tests enable advanced statistical inference from multiply imputed data.

6.4 Stepwise model selection

The standard multiple imputation scheme consists of three phases:

1. Imputation of the missing data m times;
2. Analysis of the m imputed datasets;
3. Pooling of the parameters across m analyses.

This scheme is difficult to apply if stepwise model selection is part of the statistical analysis in phase 2. Application of stepwise variable selection methods may result in sets of variables that differ across the m datasets. It is not obvious how phase 3 should be done.

6.4.1 Variable selection techniques

Brand (1999, chap. 7) was the first to recognize and treat the variable selection problem. He proposed a solution in two steps. The first step involves performing stepwise model selection separately on each imputed dataset, followed by the construction of a new supermodel that contains all variables that were present in at least half of the initial models. The idea is that this criterion excludes variables that were selected accidentally. Moreover, it is a rough correction for multiple testing. Second, a special procedure for backward elimination is applied to all variables present in the supermodel. Each variable is removed in turn, and the pooled likelihood ratio p-value (Equation 6.14) is calculated. If the largest p-value is larger than 0.05, the corresponding variable is removed, and the procedure is repeated on the smaller model. The procedure stops if all $p \leq 0.05$. The procedure was found to be a considerable improvement over complete case analysis.

Yang et al. (2005) proposed variable selection techniques using Bayesian model averaging. The authors studied two methods. The first method, called "impute then select," applies Bayesian variable selection methods on the imputed data. The second method, called "simultaneously impute and select" combines selection and missing data imputation into one Gibbs sampler. Though the latter slightly outperforms the first method, the first method is more broadly applicable. Application of the second method seems to require equivalent imputation and analysis models, thus defeating one of the main advantages of multiple imputation.

Wood et al. (2008) and Vergouwe et al. (2010) studied several scenarios for variable selection. We distinguish three general approaches:

1. *Majority*. A method that selects variables in the final that appear in at least half of the models.

2. *Stack*. Stack the imputed datasets into a single dataset, assign a fixed weight to each record and apply the usual variable selection methods.

3. *Wald*. Stepwise model selection is based on the Wald statistic calculated from the multiply imputed data.

The majority method is identical to step 1 of Brand (1999), whereas the Wald test method is similar to Brand's step 2, with the likelihood ratio test replaced by the Wald test. The Wald test method is recommended since it is a well-established approach that follows Rubin's rules, whereas the majority and stack methods fail to take into account the uncertainty caused by the missing data. Indeed, Wood et al. (2008) found that the Wald test method is the only procedure that preserved the type I error.

In practice, it may be useful to combine methods. The Wald test method is computationally intensive. `Stata` implements the Wald test approach to model selection by `mim:stepwise`. A strong point of the majority method is that it gives insight into the variability between the imputed datasets. An advantage of the stack method is that only one dataset needs to be analyzed. The discussion of Wood et al. (2008) contains additional simulations of a two-step method, in which a preselection made by the majority and stack methods is followed by the Wald test. This yielded a faster method with better theoretical properties. In practice, a judicious combination of approaches might turn out best.

6.4.2 Computation

The following steps implement a simple stepwise variable selection method in `mice`:

```
> data <- boys[boys$age >= 8, -4]
> imp <- mice(data, seed = 28382, m = 10, print = FALSE)
> expr <- expression(f1 <- lm(tv ~ 1), f2 <- step(f1,
      scope = list(upper = ~age + hgt + wgt + hc +
          gen + phb + reg), lower = ~1))
> fit <- with(imp, expr)
> formulas <- lapply(fit$analyses, formula)
> terms <- lapply(formulas, terms)
> vars <- unlist(lapply(terms, labels))
> table(vars)
```

This code imputes the `boys` data $m = 10$ times, fits a stepwise linear model to predict `tv` (testicular volume) in each of the imputed datasets and counts the number of times each variable appears in the model. The `lapply()` function is used three times. The first statement extracts the model formulas fitted to the m imputed datasets. The second `lapply()` call decomposes the model formulas into pieces, and the third call extracts the names of the variables

included in all m models. The `table()` function counts the number of times that each variable is selected as

```
> table(vars)
vars
age gen hgt phb reg wgt
 10  10   7  10  10   2
```

Variables `age`, `gen`, `phb` and (surprisingly) `reg` are always included, whereas `hgt` and `wgt` are selected 7 and 2 times, respectively. Variable `hc` is never selected, and is thus missing from the summary.

Since `hgt` appears in more than 50% of the models, we can use the Wald test to determine whether it should be in the final model.

```
> fit.without <- with(imp, lm(tv ~ age + gen + reg +
      phb))
> fit.with <- with(imp, lm(tv ~ age + gen + reg +
      phb + hgt))
> pool.compare(fit.with, fit.without)$pvalue
      [,1]
[1,] 0.297
```

The p-value is equal to 0.297, so `hgt` is not needed in the model. If we go one step further, and remove `phb`, we obtain

```
> fit.without <- with(imp, lm(tv ~ age + gen + reg))
> fit.with <- with(imp, lm(tv ~ age + gen + reg +
      phb))
> pool.compare(fit.with, fit.without)$pvalue
       [,1]
[1,] 0.0352
```

The significant difference ($p = 0.0352$) between the models implies that `phb` should be retained. We obtain similar results for the other three variables, so the final model contains `age`, `gen`, `reg` and `phb`.

6.4.3 Model optimism

The main danger of data-driven model building strategies is that the model found may depend highly on the sample at hand. For example, Viallefont et al. (2001) showed that of the variables declared to be "significant" with p-values between 0.01 and 0.05 by stepwise variable selection, only 49% actually were true risk factors. Various solutions have been proposed to counter such *model optimism*. A popular procedure is bootstrapping the model as developed in Sauerbrei and Schumacher (1992) and Harrell (2001). Although Austin (2008) found it ineffective to identify true predictors, this method has

often been found to work well for developing predictive models. The method randomly draws multiple samples with replacement from the observed sample, thus mimicking the sampling variation in the population from which the sample was drawn. Stepwise regression analyses are replayed in each bootstrap sample. The proportion of times that each prognostic variable is retained in the stepwise regression model is known as the *inclusion frequency* (Sauerbrei and Schumacher, 1992). This proportion provides information about the strength of the evidence that an indicator is an independent predictor. In addition, each bootstrap model can be fitted to the original sample. The difference between the apparent performance and the bootstrap performance provides the basis for performance measures that correct for model optimism. Steyerberg (2009, p. 95) provides an easy-to-follow procedure to calculate such *optimism-corrected performance* measures.

Clearly, the presence of missing data adds uncertainty to the model building process, so optimism can be expected to be more severe with missing data. It is not yet clear what the best way is to estimate optimism from incomplete data. Heymans et al. (2007) explored the combination of multiple imputation and the bootstrap. There appear to be at least four general procedures:

1. *Imputation*. Multiple imputation generates 100 imputed datasets. Automatic backward selection is applied to each dataset. Any differences found between the 100 fitted models are due to the missing data;

2. *Bootstrap*. 200 bootstrap samples are drawn from one singly imputed completed data. Automatic backward selection is applied to each dataset. Any differences found between the 200 fitted models are due to sampling variation;

3. *Nested Bootstrap*. The bootstrap method is applied on each of the multiply imputed datasets. Automatic backward selection is applied to each of the 100×200 datasets. Differences between the fitted model portray both sampling and missing data uncertainty;

4. *Nested Imputation*. The imputation method is applied on each of the bootstrapped datasets.

Heymans et al. (2007) observed that the imputation method produced a wider range of inclusion frequencies than the bootstrap method. This is attractive since a better separation of strong and weak predictors may ease model building. The area under the curve, or c-index, is an overall index of predictive strength. Though the type of method had a substantial effect on the apparent c-index estimate, the optimism-corrected c-index estimate was quite similar. The optimism-corrected calibration slope estimates tended to be lower in the methods involving imputation, thus necessitating more shrinkage.

Several applications of the method have now appeared (Heymans et al., 2009, 2010; Vergouw et al., 2010). Though these are promising, not all methodological questions have yet been answered (e.g., whether "nested bootstrap"

is better than "nested imputation"), so further research on this topic is necessary.

6.5 Conclusion

The statistical analysis of the multiply imputed data involved repeated analysis followed by parameter pooling. Rubin's rules apply to a wide variety of quantities, especially if these quantities are transformed toward normality. Dedicated statistical tests and model selection technique are now available. Although many techniques for complete data now have their analogues for incomplete data, the present state-of-the-art does not cover all. As multiple imputation becomes more familiar and more routine, we will see new post-imputation methodology that will be progressively more refined.

6.6 Exercises

Allison and Cicchetti (1976) investigated the interrelationship between sleep, ecological and constitutional variables. They assessed these variables for 39 mammalian species. The authors concluded that slow-wave sleep is negatively associated with a factor related to body size. This suggests that large amounts of this sleep phase are disadvantageous in large species. Also, paradoxical sleep was associated with a factor related to predatory danger, suggesting that large amounts of this sleep phase are disadvantageous in prey species.

Allison and Cicchetti (1976) performed their analyses under complete case analysis. In this exercise we will recompute the regression equations for slow wave ("nondreaming") sleep (hrs/day) and paradoxical ("dreaming") sleep (hrs/day), as reported by the authors. Furthermore, we will evaluate the imputations.

1. *Complete case analysis*. Compute the regression equations (1) and (2) from the paper of Allison and Cicchetti (1976) under complete case analysis.

2. *Imputation*. The `sleep` data are part of the `mice` package. Impute the data with `mice()` under all the default settings. Recalculate the regression equations (1) and (2) on the multiply imputed data.

3. *Traces*. Inspect the trace plot of the MICE algorithm. Does the algorithm appear to converge?

4. *More iterations.* Extend the analysis with 20 extra iterations using `mice.mids()`. Does this affect your conclusion about convergence?

5. *Distributions.* Inspect the data with diagnostic plots for univariate data. Are the univariate distributions of the observed and imputed data similar? Can you explain why they do (or do not) differ?

6. *Relations.* Inspect the data with diagnostic plots for the most interesting bivariate relations. Are the relations similar in the observed and imputed data? Can you explain why they do (or do not) differ?

7. *Defaults.* Consider each of the seven default choices from Section 5.1 in turn. Do you think the default is appropriate for your data? Explain why.

8. *Improvement.* Do you have particular suggestions for improvement? Which? Implement one (or more) of your suggestions. Do the results now look more plausible or realistic? Explain. What happened to the regression equations?

9. *Multivariate analyses.* Repeat the factor analysis and the stepwise regression. Beware: There might be pooling problems.

This page intentionally left blank

Part II

Case studies

This page intentionally left blank

Chapter 7

Measurement issues

This chapter is the first with applications to real data. A common theme is that all have "problems with the columns." Section 7.1 illustrates a number of useful steps to take when confronted with a dataset that has an overwhelming number of variables. Section 7.2 continues with the same data, and shows how a simple sensitivity analysis can be done. Section 7.3 illustrates how multiple imputation can be used to estimate overweight prevalence from self-reported data. Section 7.4 shows a way to do a sensible analysis on data that are incomparable.

7.1 Too many columns

Suppose that your colleague has become enthusiastic about multiple imputation. She asked you to create a multiply imputed version of her data, and forwarded you her entire database. As a first step, you use `R` to read it into a data frame called `data`. After this is done, you type in the following commands:

```
> library(mice)
> ## DO NOT DO THIS
> imp <- mice(data)     # not recommended
```

The program will run and impute, but after a few minutes it becomes clear that it takes a long time to finish. And after the wait is over, the imputations turn out to be surprisingly bad. What has happened?

Some exploration of the data reveals that your colleague sent you a dataset with 351 columns, essentially all the information that was sampled in the study. By default, the `mice()` function uses all other variables as predictors, so `mice()` will try to calculate regression analyses with 350 explanatory variables, and repeat that for every incomplete variable. Categorical variables are internally represented as dummy variables, so the actual number of predictors could easily double. This makes the algorithm extremely slow.

Some further exploration reveals some variables are free text fields, and that some of the missing values were not marked as such in the data. As a consequence, `mice()` treats impossible values such as "999" or "−1" as real

data. Just one forgotten missing data mark may introduce large errors into the imputations.

In order to evade such practical issues, it is necessary to spend some time exploring the data first. Furthermore, it is helpful if you understand for which scientific question the data are used. Both will help in creating sensible imputations.

This section concentrates on what can be done based on the data values themselves. In practice, it is far more productive and preferable to work together with someone who knows the data really well, and who knows the questions of scientific interest that one could ask from the data. Sometimes the possibilities for cooperation can be limited. This may occur, for example, if the data have come from several external sources (as in meta analysis), or if the dataset is so diverse that no one person can cover all of its contents. It will be clear that this situation calls for a careful assessment of the data quality, well before attempting imputation.

7.1.1 Scientific question

There is a paradoxical inverse relation between blood pressure (BP) and mortality in persons over 85 years of age (Boshuizen et al., 1998; Van Bemmel et al., 2006). Normally, people with a lower BP live longer, but the oldest old with lower BP live a shorter time.

The goal of the study was to determine if the relation between BP and mortality in the very old is due to frailty. A second goal was to know whether high BP was a still risk factor for mortality after the effects of poor health had been taken into account.

The study compared two models:

1. The relation between mortality and BP adjusted for age, sex and type of residence.

2. The relation between mortality and BP adjusted for age, sex, type of residence and health.

Health was measured by 28 different variables, including mental state, handicaps, being dependent in activities of daily living, history of cancer and others. Including health as a set of covariates in model 2 might explain the relation between mortality and BP, which, in turn, has implications for the treatment of hypertension in the very old.

7.1.2 Leiden 85+ Cohort

The data come from the 1236 citizens of Leiden who were 85 years or older on December 1, 1986 (Lagaay et al., 1992; Izaks et al., 1997). These individuals were visited by a physician between January 1987 and May 1989. A full medical history, information on current use of drugs, a venous blood sample, and

other health-related data were obtained. BP was routinely measured during the visit. Apart from some individuals who were bedridden, BP was measured while seated. An Hg manometer was used and BP was rounded to the nearest 5 mmHg. Measurements were usually taken near the end of the interview. The mortality status of each individual on March 1, 1994 was retrieved from administrative sources.

Of the original cohort, a total of 218 persons died before they could be visited, 59 persons did not want to participate (some because of health problems), 2 emigrated and 1 was erroneously not interviewed, so 956 individuals were visited. Effects due to subsampling the visited persons from the entire cohort were taken into account by defining the date of the home visit as the start (Boshuizen et al., 1998). This type of selection will not be considered further.

7.1.3 Data exploration

The data are stored as a `SAS` export file. The `read.xport()` function from the `foreign` package can read the data.

```
> library(foreign)
> file.sas <- file.path(project, "original/master85.xport")
> original.sas <- read.xport(file.sas)
> names(original.sas) <- tolower(names(original.sas))
> dim(original.sas)
[1] 1236  351
```

The dataset contains 1236 rows and 351 columns. When I tracked down the origin of the data, the former investigators informed me that the file was composed during the early 1990s from several parts. The basic component consisted of a `Dbase` file with many free text fields. A dedicated `Fortran` program was used to separate free text fields. All fields with medical and drug-related information were hand-checked against the original forms. The information not needed for analysis was not cleaned. All information was kept, so the file contains several versions of the same variable.

A first scan of the data makes clear that some variables are free text fields, person codes and so on. Since these fields cannot be sensibly imputed, they are removed from the data. In addition, only the 956 cases that were initially visited are selected, as follows:

```
> all <- names(original.sas)
> drop <- c(3, 22, 58, 162:170, 206:208)
> keep <- !(1:length(all) %in% drop)
> leiden85 <- original.sas[original.sas$abr == "1",
      keep]
> data <- leiden85
```

The frequency distribution of the missing cases per variable can be obtained as:

```
> ini <- mice(data, maxit=0)   # recommended
> table(ini$nmis)
  0   2   3   5   7  14  15  28  29  32  33  34  35  36  40
 87   2   1   1   1   1   2   1   3   2  34  15  25   4   1
 42  43  44  45  46  47  48  49  50  51  54  64  72  85 103
  1   2   1   4   2   3  24   4   1  20   2   1   4   1   1
121 126 137 155 157 168 169 201 202 228 229 230 231 232 233
  1   1   1   1   1   2   1   7   3   5   4   2   4   1   1
238 333 350 501 606 635 636 639 642 722 752 753 812 827 831
  1   3   1   3   1   2   1   1   2   1   5   3   1   1   3
880 891 911 913 919 928 953 954 955
  3   3   3   1   1   1   3   3   3
```

Thus, there are 87 variables that are complete. The set includes administrative variables (e.g., person number), design factors, date of measurement, survival indicators, selection variables and so on. The set also included some variables for which the missing data were inadvertently not marked, containing values such as "999" or "−1." For example, the frequency distribution of the complete variable 'beroep1' (occupation) is

```
> table(data$beroep1, useNA = "always")
  -1    0    1    2    3    4    5    6 <NA>
  42    1  576  125  104   47   44   17    0
```

There are no missing values, but a variable with just categories "−1" and "0" is suspect. The category "−1" likely indicates that the information was missing (this was the case indeed). One option is to leave this "as is," so that `mice()` treats it as complete information. All cases with a missing occupation are then seen as a homogeneous group.

Two other variables without missing data markers are `syst` and `diast`, i.e., systolic and diastolic BP classified into six groups. The correlation (using the observed pairs) between `syst` and `rrsyst`, the variable of primary interest, is 0.97. Including `syst` into the imputation model for `rrsyst` will ruin the imputations. The "as is" option is dangerous, and shares some of the same perils of the indicator method (cf. Section 1.3.7). The message is that variables that are 100% complete deserve appropriate attention.

After a first round of screening, I found that 57 of the 87 complete variables were uninteresting or problematic in some sense. Their names were placed on a list named `outlist1` as follows:

```
> v1 <- names(ini$nmis[ini$nmis == 0])
> outlist1 <- v1[c(1, 3:5, 7:10, 16:47, 51:60, 62,
      64:65, 69:72)]
> length(outlist1)
[1] 57
```

7.1.4 Outflux

We should also scrutinize the variables at the other end. Variables with high proportions of missing data generally create more problems than they solve. Unless some of these variables are of genuine interest to the investigator, it is best to leave them out. Virtually every dataset contains some parts that could better be removed before imputation. This includes, but is not limited to, uninteresting variables with a high proportion of missing data, variables without a code for the missing data, administrative variables, constant variables, duplicated, recoded or standardized variables, and aggregates and indices of other information.

Figure 7.1 is the influx-outflux pattern of Leiden 85+ Cohort data. The influx of a variable quantifies how well its missing data connect to the observed data on other variables. The outflux of a variable quantifies how well its observed data connect to the missing data on other variables. See Section 4.1.3 for more details. Though the display could obviously benefit from a better label-placing strategy, we can see three groups. All points are relatively close to the diagonal, which indicates that influx and outflux are balanced.

The group at the left-upper corner has (almost) complete information, so the number of missing data problems for this group is relatively small. The intermediate group has an outflux between 0.5 and 0.8, which is small. Missing data problems are more severe, but potentially this group could contain important variables. The third group has an outflux with 0.5 and lower, so its predictive power is limited. Also, this group has a high influx, and is thus highly dependent on the imputation model.

Note that there are two variables (`hypert1` and `aovar`) in the third group that are located above the diagonal. Closer inspection reveals that the missing data mark had not been set for these two variables. Variables that might cause problems later on in the imputations are located in the lower-right corner. Under the assumption that this group does not contain variables of scientific interest, I transferred 45 variables with an outflux < 0.5 to `outlist2`:

```
> outlist2 <- row.names(fx)[fx$outflux < 0.5]
> length(outlist2)
[1] 45
```

In these data, the set of selected variables is identical to the group with more than 500 missing values, but this need not always be the case. I removed

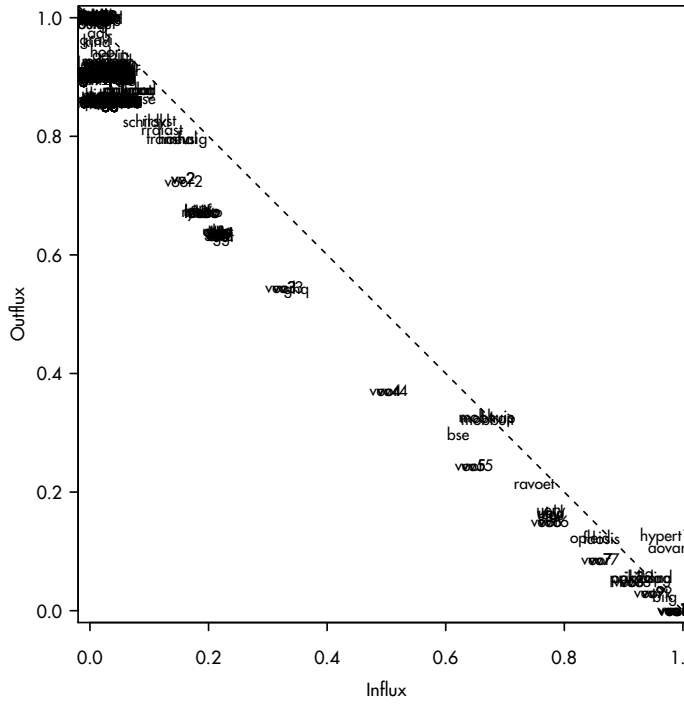


Figure 7.1: Global influx-outflux pattern of the Leiden 85+ Cohort data. Variables with higher outflux are (potentially) the more powerful predictors. Variables with higher influx depend stronger on the imputation model.

the 45 variables, recalculated influx and outflux on the smaller dataset and selected 32 new variables with outflux < 0.5.

```
> data2 <- data[, !names(data) %in% outlist2]
> fx2 <- flux(data2)
> outlist3 <- row.names(fx2)[fx2$outflux < 0.5]
```

Variable `outlist3` contains 32 variable names, among which are many laboratory measurements. I prefer to keep these for imputation since they may correlate well with BP and survival. Note that the outflux changed considerably as I removed the 45 least observed variables. Influx remained nearly the same.

7.1.5 Logged events

Another source of information is a list of logged events produced by `mice()`. The logged events are a structured report that identify problems

with the data, as well as corrective actions taken by `mice()`. It is a component called `loggedEvents` of the `mids` object.

```
> ini$log[1:3, ]
  it im co dep     meth out
1  0  0  0     constant abr
2  0  0  0     constant vo7
3  0  0  0     constant vo9
```

The log contains variables with a problem. `mice()` silently removes problematic variables, but leaves an entry in the log. At initialization, a log entry is made for the following actions:

- A variable that contains missing values, that is not imputed and that is used as a predictor is removed;
- A constant variable is removed;
- A collinear variable is removed.

During execution of the algorithm log entries signal the following actions:

- One or more variables that are linearly dependent are removed;
- Proportional odds imputation did not converge and was replaced by the multinomial model.

The log is a data frame with six columns. The columns `it`, `im` and `co` stand for iteration, imputation number and column number, respectively. The column `dep` contains the name of the active variable. The column `meth` entry signals the type of problem. Finally, the column `out` contains the name(s) of the removed variable(s). The log contains valuable information about methods and variables that were difficult to fit. Closer examination of the data might then reveal what the problem is.

Based on the initial analysis by `mice()`, I placed the names of all constant and collinear variables on `outlist4` by

```
> outlist4 <- as.character(ini$log[, "out"])
```

The outlist contains 28 variables.

7.1.6 Quick predictor selection for wide data

The `mice` package contains the function `quickpred()`, which implements the predictor selection strategy of Section 5.3.2. In order to apply this strategy to the Leiden 85+ Cohort data, I first deleted the variables on three of the four outlists created in the previous sections.

```
> outlist <- unique(c(outlist1, outlist2, outlist4))
> length(outlist)
[1] 108
```

There are 108 unique variables to be removed. Thus, before doing any imputations, I cleaned out about one third of the data that are likely to cause problems. The downsized data are

```
> data2 <- data[, !names(data) %in% outlist]
```

The next step is to build the imputation model according to the strategy outlined above. The function `quickpred()` is applied as follows:

```
> inlist <- c("sex", "lftanam", "rrsyst", "rrdiast")
> pred <- quickpred(data2, minpuc = 0.5, inc = inlist)
```

There are 198 incomplete variables in `data2`. The character vector `inlist` specifies the names of the variables that should be included as covariates in every imputation model. Here I specified age, sex and blood pressure. Blood pressure is the variable of central interest, so I included it in all models. This list could be longer if there are more outcome variables. The `inlist` could also include design factors.

The `quickpred()` function creates a binary predictor matrix of 198 rows and 198 columns. The rows correspond to the incomplete variables and the columns report the same variables in their role as predictor. The number of predictors varies per row. We can display the distribution of the number of predictors by

```
> table(rowSums(pred))
 0  7 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
30  1  2  1  1  2  5  2 13  8 16  9 13  7  5  6 10  6  3  6
29 30 31 32 33 34 35 36 37 38 39 40 41 42 44 45 46 49 50 57
 4  8  3  6  9  2  4  6  2  5  2  4  2  3  4  3  3  3  1  1
59 60 61 68 79 83 85
 1  1  1  1  1  1  1
```

The variability in model sizes is substantial. The 30 rows with no predictors are complete. The mean number of predictors is equal to 24.8. It is possible to influence the number of predictors by altering the values of `mincor` and `minpuc` in `quickpred()`. A number of predictors of 15–25 is about right (cf. Section 5.3.2), so I decided to accept this predictor matrix. The number of predictors for systolic and diastolic BP are

```
> rowSums(pred[c("rrsyst", "rrdiast"), ])
 rrsyst rrdiast
     41      36
```

The names of the predictors for `rrsyst` can be obtained by

```
> names(data2)[pred["rrsyst", ] == 1]
```

It is sometimes useful the inspect the correlations of the predictors selected by `quickpred()`. Table 3 in Van Buuren et al. (1999) provides an example. For a given variable, the correlations can be tabulated by

```
> vname <- "rrsyst"
> y <- cbind(data2[vname], r = !is.na(data2[, vname]))
> vdata <- data2[, pred[vname, ] == 1]
> round(cor(y = y, x = vdata, use = "pair"), 2)
```

7.1.7 Generating the imputations

Everything is now ready to impute the data as

```
> imp.qp <- mice(data2, pred = pred, seed = 29725)
```

Thanks to the smaller dataset and the more compact imputation model, this code runs about 50 times faster than "blind imputation" as practiced in Section 7.1. More importantly, the new solution is much better. To illustrate the latter, take a look at Figure 7.2.

The figure is the scatterplot of `rrsyst` and `rrdiast` of the first imputed dataset. The left-hand figure shows what can happen if the data are not properly screened. In this particular instance, a forgotten missing data mark of "−1" was counted as a valid blood pressure value, and produced imputation that are far off. In contrast, the imputations created with the help of `quickpred()` look reasonable.

The plot was created by the following code:

```
> vnames <- c("rrsyst", "rrdiast")
> cd1 <- complete(imp)[, vnames]
> cd2 <- complete(imp.qp)[, vnames]
> typ <- factor(rep(c("blind imputation", "quickpred"),
      each = nrow(cd1)))
> mis <- ici(data2[, vnames])
> mis <- is.na(imp$data$rrsyst) | is.na(imp$data$rrdiast)
> cd <- data.frame(typ = typ, mis = mis, rbind(cd1,
      cd2))
> xyplot(jitter(rrdiast, 10) ~ jitter(rrsyst, 10) |
      typ, data = cd, groups = mis, col = c(mdc(1),
      mdc(2)), xlab = "Systolic BP (mmHg)", type = c("g",
      "p"), ylab = "Diastolic BP (mmHg)", pch = c(1,
      19), strip = strip.custom(bg = "gray95"),
      scales = list(alternating = 1, tck = c(1,
          0)))
```

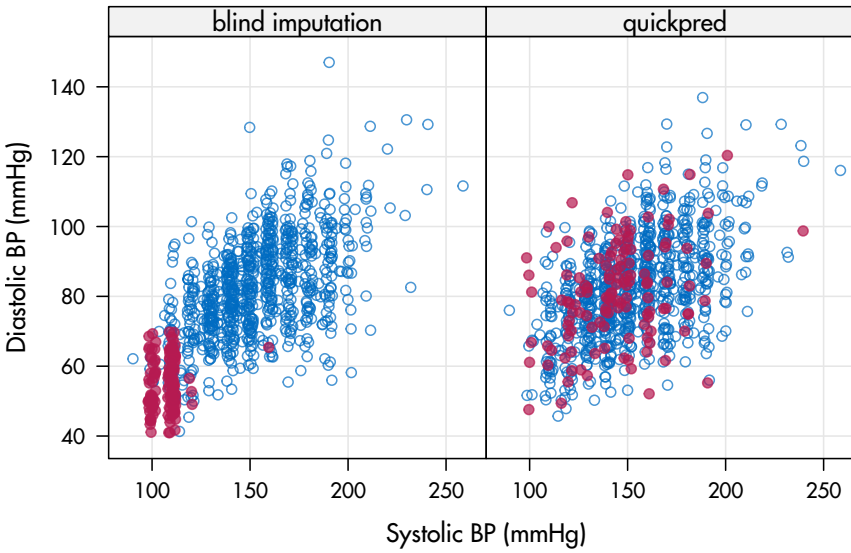


Figure 7.2: Scatterplot of systolic and diastolic blood pressure from the first imputation. The left-hand-side plot was obtained after just running `mice()` on the data without any data screening. The right-hand-side plot is the result after cleaning the data and setting up the predictor matrix with `quickpred()`. Leiden 85+ Cohort data.

7.1.8 A further improvement: Survival as predictor variable

If the complete-data model is a survival model, incorporating the cumulative hazard to the survival time, $H_0(T)$, as one of the predictors provide slightly better imputations (White and Royston, 2009). In addition, the event indicator should be included into the model. The Nelson-Aalen estimate of $H_0(T)$ in the Leiden 85+ Cohort can be calculated as

```
> dat <- cbind(data2, dead = 1 - data2$dwa)
> hazard <- nelsonaalen(dat, survda, dead)
```

where `dead` is coded such that "1" means death. The `nelsonaalen()` function is part of `mice`. Table 7.1 lists the correlations beween several key variables. The correlation between $H_0(T)$ and T is almost equal to 1, so for these data it matters little whether we take $H_0(T)$ or T as the predictor. The high correlation may be caused by the fact that nearly everyone in this cohort has died, so the percentage of censoring is low. The correlation between $H_0(T)$ and T could be lower in other epidemiological studies, and thus it might matter whether we take $H_0(T)$ or T. Observe that the correlation between $\log(T)$ and blood pressure is higher than for $H_0(T)$ or T, so it makes sense to add $\log(T)$ as an

Table 7.1: Pearson correlations between the cumulative death hazard $H_0(T)$, survival time T, $\log(T)$, systolic and diastolic blood pressure.

	$H_0(T)$	T	$\log(T)$	SBP	DBP
$H_0(T)$	1.000	0.997	0.830	0.169	0.137
T	0.997	1.000	0.862	0.176	0.141
$\log(T)$	0.830	0.862	1.000	0.205	0.151
SBP	0.169	0.176	0.205	1.000	0.592
DBP	0.137	0.141	0.151	0.592	1.000

additional predictor. This strong relation may have been a consequence of the design, as the frail people were measured first.

7.1.9 Some guidance

Imputing data with many columns is challenging. Even the most carefully designed and well-maintained data may contain information or errors that can send the imputations awry. I conclude this section by summarizing advice for imputation of data with "too many columns."

1. Inspect all complete variables for forgotten missing data marks. Repair or remove these variables. Even one forgotten mark may ruin the imputation model. Remove outliers with improbable values.

2. Obtain insight into the strong and weak parts of the data by studying the influx-outflux pattern. Unless they are scientifically important, remove variables with low outflux, or with high fractions of missing data.

3. Perform a dry run with `maxit=0` and inspect the logged events produced by `mice()`. Remove any constant and collinear variables before imputation.

4. Find out what will happen after the data have been imputed. Determine a set of variables that are important in subsequent analyses, and include these as predictors in all models. Transform variables to improve predictability and coherence in the complete-data model.

5. Run `quickpred()`, and determine values of `mincor` and `minpuc` such that the average number of predictors is around 25.

6. After imputation, determine whether the generated imputations are sensible by comparing them to the observed information, and to knowledge external to the data. Revise the model where needed.

7. Document your actions and decisions, and obtain feedback from the owner of the data.

It is most helpful to try out these techniques on data gathered within your own institute. Some of these steps may not be relevant for other data. Determine where you need to adapt the procedure to suit your needs.

7.2 Sensitivity analysis

The imputations created in Section 7.1 are based on the assumption that the data are MAR (cf. Sections 1.2 and 2.2.4). While this is often a good starting assumption, it may not be realistic for the data at hand. When the data are not MAR, we can follow two strategies to obtain plausible imputations. The first strategy is to make the data "more MAR." In particular, this strategy requires us to identify additional information that explains differences in the probability to be missing. This information is then used to generate imputations conditional on that information. The second strategy is to perform a sensitivity analysis. The goal of the sensitivity analysis is to explore the result of the analysis under alternative scenarios for the missing data. See Section 5.2 for a more elaborate discussion of these strategies.

This section explores sensitivity analysis for the Leiden 85+ Cohort data. In sensitivity analysis, imputations are generated according to one or more scenarios. The number of possible scenarios is infinite, but these are not equally likely. A scenario could be very simple, like assuming that everyone with a missing value had scored a "yes," or assuming that those with missing blood pressures have the minimum possible value. While easy to interpret, such extreme scenarios are highly unlikely. Preferably, we should attempt to make an educated guess about both the direction and the magnitude of the missing data had they been observed. By definition, this guess needs to be based on external information beyond the data.

7.2.1 Causes and consequences of missing data

We continue with the Leiden 85+ Cohort data described in Section 7.1. The objective is to estimate the effect of blood pressure (BP) on mortality. BP was not measured for 126 individuals (121 systolic, 126 diastolic).

The missingness is strongly related to survival. Figure 7.3 displays the Kaplan-Meier survival curves for those with ($n = 835$) and without ($n = 121$) a measurement of systolic BP (SBP). BP measurement was missing for a variety of reasons. Sometimes there was a time constraint. In other cases the investigator did not want to place an additional burden on the respondent. Some subjects were too ill to be measured.

Table 7.2 indicates that BP was measured less frequently for very old persons and for persons with health problems. Also, BP was measured more often if the BP was too high, for example if the respondent indicated a previous

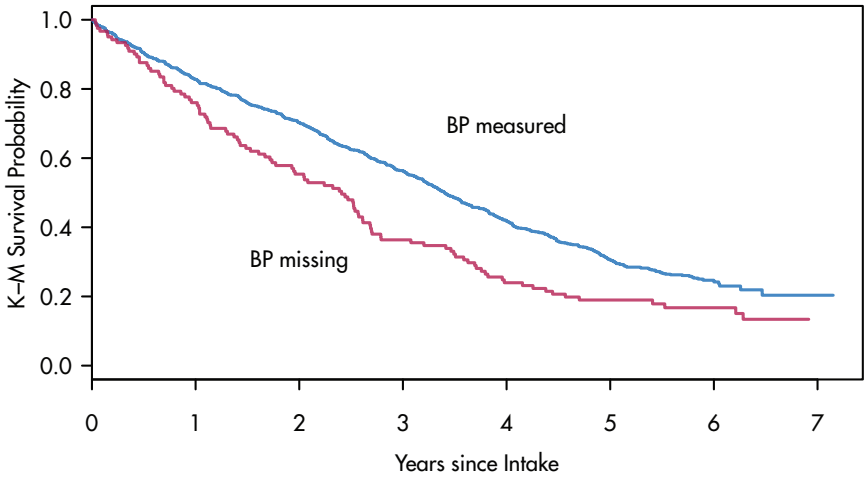


Figure 7.3: Kaplan–Meier curves of the Leiden 85+ Cohort, stratified according to missingness. The figure shows the survival probability since intake for the group with observed BP measures (blue) and the group with missing BP measures (red).

Table 7.2: Some variables that have different distributions in the response ($n = 835$) and non-response groups ($n = 121$). Shown are rounded percentages. Significance levels correspond to the χ^2-test.

Variable		Observed BP	Missing BP
Age (year)	$p < 0.0001$		
85–89		63	48
90–94		32	34
95+		6	18
Type of residence	$p < 0.0001$		
Independent		52	35
Home for elderly		35	54
Nursing home		13	12
Activities of daily living (ADL)	$p < 0.001$		
Independent		73	54
Dependent on help		27	46
History of hypertension	$p = 0.06$		
No		77	85
Yes		23	15

Table 7.3: Proportion of persons for which no BP was measured, cross-classified by three-year survival and previous hypertension history. Shown are proportions per cell (number of cases with missing BP/total cell count).

Survived	History of Previous Hypertension			
	No		Yes	
Yes	8.7%	(34/390)	8.1%	(10/124)
No	19.2%	(69/360)	9.8%	(8/82)

diagnosis of hypertension, or if the respondent used any medication against hypertension. The missing data rate of BP also varied during the period of data collection. The rate gradually increases during the first seven months of the sampling period from 5 to 40 percent of the cases, and then suddenly drops to a fairly constant level of 10–15 percent. A complicating factor here is that the sequence in which the respondents were interviewed was not random. High-risk groups, that is, elderly in hospitals and nursing homes and those over 95, were visited first.

Table 7.3 contains the proportion of persons for which BP was not measured, cross-classified by three-year survival and history of hypertension as measured during anamnesis. Of all persons who die within three years and that have no history of hypertension, more than 19% have no BP score. The rate for other categories is about 9%. This suggests that a relatively large group of individuals without hypertension and with high mortality risk is missing from the sample for which BP is known.

Using only the complete cases could lead to confounding by selection. The complete case analysis might underestimate the mortality of the lower and normal BP groups, thereby yielding a distorted impression of the influence of BP on survival. This reasoning is somewhat tentative as it relies on the use of hypertension history as a proxy for BP. If true, however, we would expect more missing data from the lower BP measures. It is known that BP and mortality are inversely related in this age group, that is, lower BP is associated with higher mortality. If there are more missing data for those with low BP and high mortality (as in Table 7.3), selection of the complete cases could blur the effect of BP on mortality.

7.2.2 Scenarios

The previous section presented evidence that there might be more missing data for the lower blood pressures. Imputing the data under MAR can only account for nonresponse that is related to the observed data. However, the missing data may also be caused by factors that have not been observed. In order to study the influence of such factors on the final inferences, let us conduct a sensitivity analysis.

Section 3.9 advocated the use of simple adjustments to the imputed data as

Table 7.4: Realized difference in means of the observed and imputed SBP (mmHg) data under various δ-adjustments. The number of multiple imputations is $m = 5$.

δ	Difference
0	−8.2
−5	−12.3
−10	−20.7
−15	−26.1
−20	−31.5

a way to perform sensitivity analysis. Table 3.6 lists possible values for an offset δ, together with an interpretation whether the value would be (too) small or (too) large. The next section uses the following range for δ: 0 mmHg (MCAR, too small), −5 mmHg (small), −10 mmHg (large), −15 mmHg (extreme) and −20 mmHg (too extreme). The last value is unrealistically low, and is primarily included to study the stability of the analysis in the extreme.

7.2.3 Generating imputations under the δ-adjustment

Subtracting a fixed amount from the imputed values is easily achieved by the `post` processing facility in `mice()`. The following code first imputes under $\delta = 0$ mmHg (MAR), then under $\delta = -5$ mmHg, and so on.

```
> delta <- c(0, -5, -10, -15, -20)
> post <- imp.qp$post
> imp.all.undamped <- vector("list", length(delta))
> for (i in 1:length(delta)) {
      d <- delta[i]
      cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +",
          d)
      post["rrsyst"] <- cmd
      imp <- mice(data2, pred = pred, post = post,
          maxit = 10, seed = i * 22)
      imp.all.undamped[[i]] <- imp
  }
```

Note that we specify an adjustment in SBP only. Since imputed SBP is used to impute other incomplete variables, δ will also affect the imputations in those. The strength of the effect depends on the correlation between SBP and the variable. Thus, using a δ-adjustment for just one variable will affect many.

The mean of the observed systolic blood pressures is equal to 152.9 mmHg. Table 7.4 provides the differences in means between the imputed and observed data as a function of δ. For $\delta = 0$, i.e., under MAR, we find that the imputations are on average 8.2 mmHg lower than the observed blood pressure, which

is in line with the expectations. As intended, the gap between observed and imputed increases as δ decreases.

Note that for $\delta = -10$ mmHg, the magnitude of the difference with the MAR case ($-20.7+8.2 = -12.5$ mmHg) is somewhat larger in size than δ. The same holds for $\delta = -15$ mmHg and $\delta = -20$ mmHg. This is due to feedback of the δ-adjustment itself via third variables. It is possible to correct for this, for example by multiplying δ by a damping factor $\sqrt{1-r^2}$, with r^2 the proportion of explained variance of the imputation model for SBP. In `R` this can be done by changing the expression for `cmd` as

```
> cmd <- paste("fit <- lm(y ~ as.matrix(x));
                damp <- sqrt(1 - summary(fit)$r.squared);
                imp[[j]][, i] <- imp[[j]][, i] + damp * ", d)
```

As the results of the complete-data analysis turned out to be very similar to the "raw" δ, this route is not further explored.

7.2.4 Complete data analysis

Complete data analysis is a Cox regression with survival since intake as the outcome, and with blood pressure groups as the main explanatory variable. The analysis is stratified by sex and age group. The preliminary data transformations needed for this analysis were performed as follows:

```
> cda <- expression(
      sbpgp <- cut(rrsyst, breaks = c(50, 124,
            144, 164, 184, 200, 500)),
      agegp <- cut(lftanam, breaks = c(85, 90,
            95, 110)),
      dead  <- 1 - dwa,
      coxph(Surv(survda, dead)
            ~ C(sbpgp, contr.treatment(6, base = 3))
            + strata(sexe, agegp)))
> imp <- imp.all.damped[[1]]
> fit <- with(imp, cda)
```

The `cda` object is an expression vector containing several statements needed for the complete data analysis. The `cda` object will be evaluated within the environment of the imputed data, so (imputed) variables like `rrsyst` and `survda` are available during execution. Derived variables like `sbpgp` and `agegp` are temporary and disappear automatically. When evaluated, the expression vector returns the value of the last expression, in this case the object produced by `coxph()`. The expression vector provides a flexible way to apply `R` code to the imputed data. Do not forget to include commas to separate the individual expressions. The pooled hazard ratio per SBP group can be calculated by

```
> as.vector(exp(summary(pool(fit))[, 1]))
```

Table 7.5: Hazard ratio estimates (with 95% confidence interval) of the classic proportional hazards model. The estimates are relative to the reference group (145–160 mmHg). Rows correspond to different scenarios in the δ-adjustment. The row labeled "CCA" contains results of the complete case analysis.

δ	<125 mmHg		125–140 mmHg		>200 mmHg	
0	1.76	(1.36–2.28)	1.43	(1.16–1.77)	0.86	(0.44–1.67)
-5	1.81	(1.42–2.30)	1.45	(1.18–1.79)	0.88	(0.50–1.55)
-10	1.89	(1.47–2.44)	1.50	(1.21–1.86)	0.90	(0.51–1.59)
-15	1.82	(1.39–2.40)	1.45	(1.14–1.83)	0.88	(0.49–1.57)
-20	1.80	(1.39–2.35)	1.46	(1.17–1.83)	0.85	(0.48–1.50)
CCA	1.76	(1.36–2.28)	1.48	(1.19–1.84)	0.89	(0.51–1.57)

```
[1] 1.758 1.433 1.065 1.108 0.861
```

Table 7.5 provides the hazard ratio estimates under the different scenarios for three SBP groups. A risk ratio of 1.76 means that the mortality risk (after correction for sex and age) in the group "SBP <125 mmHg" is 1.76 times the risk of the reference group "145–160 mmHg." The inverse relation relation between mortality and blood pressure in this age group is consistent, where even the group with the highest blood pressures have (nonsignificant) lower risks.

Though the imputations differ dramatically under the various scenarios, the hazard ratio estimates for different δ are close. Thus, the results are essentially the same under all specified MNAR mechanisms. Also observe that the results are close to those from the analysis of the complete cases.

7.2.5 Conclusion

Sensitivity analysis is an important tool for investigating the plausibility of the MAR assumption. This section explored the use of an informal, simple and direct method to create imputations under nonignorable models by simply deducting some amount from the imputations.

Section 3.9.1 discussed shift, scale and shape parameters for nonignorable models. We only used a shift parameter here, which suited our purposes in the light of what we knew about the causes of the missing data. In other applications, scale or shape parameters could be more natural. The calculations are easily adapted to such cases.

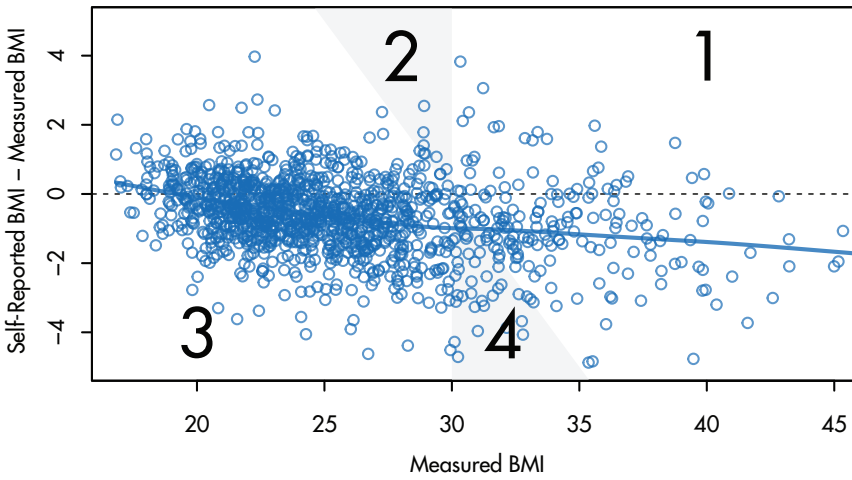


Figure 7.4: Underestimation of obesity prevalence in self-reported data. Self-reported BMI is on average 1–2 kg/m^2 too low. Lines are fitted by lowess.

7.3 Correct prevalence estimates from self-reported data

7.3.1 Description of the problem

Prevalence estimates for overweight and obesity are preferably based on standardized measured data of height and weight. However, obtaining such measures is logistically challenging and costly. An alternative is to ask persons to report their own height and weight. It is well known that such measures are subject to systematic biases. People tend to overestimate their height and underestimate their weight. A recent overview covering 64 studies can be found in Gorber et al. (2007).

Body Mass Index (BMI) is calculated from height and weight as kg/m^2. For BMI both biases operate in the same direction, so any self-reporting biases are amplified in BMI. Figure 7.4 is drawn from data of Krul et al. (2010). Self-reported BMI is on average 1–2 kg/m^2 lower than measured BMI.

BMI values can be categorized into underweight (BMI $<$ 18.5), normal (18.5 $\leq$ BMI $<$ 25), overweight (25 $\leq$ BMI $<$ 30), and obese (BMI $\geq$ 30). Self-reported BMI may assign subjects to a category that is too low. In Figure 7.4 persons in the white area labeled "1" are obese according to both self-reported and measured BMI. Persons in the white area labeled "3" are non-obese. The shaded areas represent disagreement between measured and self-reported obesity. The shaded area "4" are obese according to measured BMI, but not to self-report. The reverse holds for the shaded area "2." Due

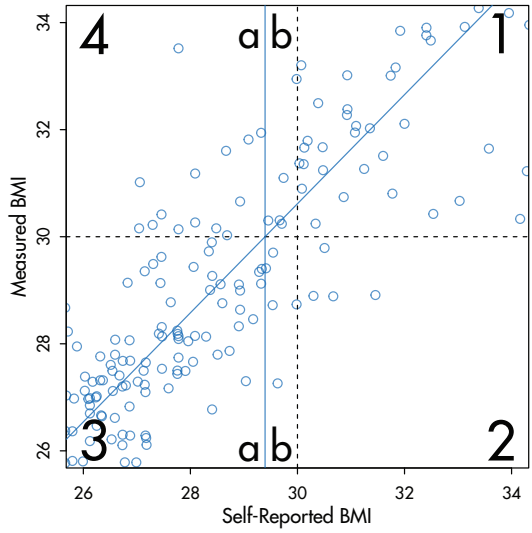


Figure 7.5: Illustration of the bias of predictive equations. In general, the combined region 2 + 3b will have fewer cases than region 4a. This causes a downward bias in the prevalence estimate.

to self-reporting bias, the number of persons located in area "4" is generally larger than in area "2," leading to underestimation.

There have been many attempts to correct measured height and weight for bias using predictive equations. These attempts have generally not been successful. The estimated prevalences were often still found to be too low after correction. Moreover, there is substantial heterogeneity in the proposed predictive formulae, resulting in widely varying prevalence estimates. See Visscher et al. (2006) for a summary of these issues. The current consensus is that it is not possible to estimate overweight and obesity prevalence from self-reported data. Dauphinot et al. (2008) even suggested to lower cut-off values for obesity based on self-reported data.

The goal is to estimate obesity prevalence in the population from self-reported data. This estimate should be unbiased in the sense that, on average, it should be equal to the estimate that would have been obtained had data been truly measured. Moreover, the estimate must be accompanied by a standard error or a confidence interval.

7.3.2 Don't count on predictions

Table 4 in Visscher et al. (2006) lists 36 predictive equations that have been proposed over the years. Visscher et al. (2006) observed that these equations predict too low. This section explains why this happens.

Figure 7.5 plots the data of Figure 7.4 in a different way. The figure is centered around the BMI of 30 kg/m^2. The two dashed lines divide the area into four quadrants. Quadrant 1 contains the cases that are obese according to both BMI values. Quadrant 3 contains the cases that are classified as non-obese according to both. Quadrant 2 holds the subjects that are classified as obese according to self-report, but not according to measured BMI. Quadrant 4 has the opposite interpretation. The area and quadrant numbers used in Figures 7.4 and 7.5 correspond to identical subdivisions in the data.

The "true obese" in Figure 7.5 lie in quadrants 1 and 4. The obese according to self-report are located in quadrants 1 and 2. Observe that the number of cases in quadrant 2 is smaller than in quadrant 4, a result of the systematic bias that is observed in humans. Using uncorrected self-report thus leads to an underestimate of the true prevalence.

The regression line that predicts measured BMI from self-reported BMI is added to the display. This line intersects the horizontal line that separates quadrant 3 from quadrant 4 at a (self-reported) BMI value of 29.4 kg/m^2. Note that using the regression line to predict obese versus non-obese is in fact equivalent to classifying all cases with a self-report of 29.4 kg/m^2 or higher as obese. Thus, the use of the regression line as a predictive equation effectively shifts the vertical dashed line from 30 kg/m^2 to 29.4 kg/m^2. Now we can make the same type of comparison as before. We count the number of cases in quadrant 2 + section 3b (n_1), and compare it to the count in region 4a (n_2). The difference $n_2 - n_1$ is now much smaller, thanks to the correction by the predictive equation.

However, there is still bias remaining. This comes from the fact that the distribution on the left side is more dense. The number of subjects with a BMI of 28 kg/m^2 is typically larger than the number of subjects with a BMI of 32 kg/m^2. Thus, even if a symmetric normal distribution around the regression line is correct, n_2 is on average larger than n_1. This yields bias in the predictive equation.

Observe that this effect will be stronger if the regression line becomes more shallow, or equivalently, if the spread around the regression line increases. Both are manifestation of less-than-perfect predictability. Thus, predictive equations only work well if the predictability is very high, but they are systematically biased in general.

7.3.3 The main idea

Table 7.6 lists the six variable names needed in this application. Let us assume that we have two data sources available:

- The *calibration dataset* contains n_c subjects for which both self-reported and measured data are available;
- The *survey dataset* contains n_s subjects with only the self-reported data.

Table 7.6: Basic variables needed to correct overweight/obesity prevalence for self-reporting.

Name	Description
`age`	Age (years)
`sex`	Sex (M/F)
`hm`	Height measured (cm)
`hr`	Height reported (cm)
`wm`	Weight measured (kg)
`wr`	Weight reported (kg)

Note: The survey data are representative for the population of interest, possibly after correction for design factors.

We assume that the common variables in these two datasets are comparable.

The idea is to stack the datasets, multiply impute the missing values for `hm` and `wm` in the survey data and estimate the overweight and obesity prevalence (and their standard errors) from the imputed survey data. See Schenker et al. (2010) for more background.

7.3.4 Data

The calibration sample is taken from Krul et al. (2010). The dataset contains of $n_c = 1257$ Dutch subjects with both measured and self-reported data. The survey sample consists of $n_s = 803$ subjects of a representative sample of Dutch adults aged 18–75 years. These data were collected in November 2007 either online or using paper-and-pencil methods. The missing data pattern in the combined data is summarized as

```
> md.pattern(selfreport[, c("age", "sex", "hm",
      "hr", "wm", "wr")])
     age sex hr wr  hm  wm
1257   1   1  1  1   1   1    0
 803   1   1  1  1   0   0    2
       0   0  0  0 803 803 1606
```

The row containing all ones corresponds to the 1257 observations from the calibration sample with complete data, whereas the rows with a zero on `hm` and `wm` correspond to 803 observations from the survey sample (where `hm` and `wm` were not measured).

We apply predictive mean matching (cf. Section 3.4) to impute `hm` and `wm` in the 803 records from the survey data. The number of imputations $m = 10$. The complete-data estimates are calculated on each imputed dataset and combined using Rubin's pooling rules to obtain prevalence rates and the associated confidence intervals as in Sections 2.3.2 and 2.4.

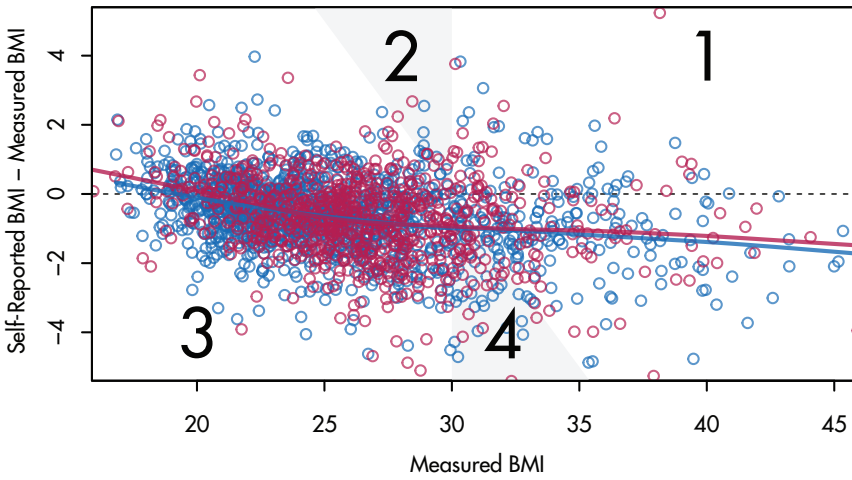


Figure 7.6: Relation between measured BMI and self-reported BMI in the calibration (blue) and survey (red) data in the first imputed dataset.

7.3.5 Application

The `mice()` function can be used to create $m = 10$ multiply imputed datasets. We imputed measured height, measured weight and and measured BMI using the following code:

```
> bmi <- function(h, w) {
      return(w/(h/100)^2)
  }
> init <- mice(selfreport, maxit = 0)
> meth <- c(rep("", 5), "pmm", "pmm", rep("", 6),
      "~bmi(hm,wm)", "")
> pred <- init$pred
> pred[, c("src", "id", "pop", "prg", "edu", "etn",
      "web", "bm", "br")] <- 0
> imp <- mice(selfreport, pred = pred, meth = meth,
      seed = 66573, maxit = 20, m = 10)
```

The code defines a `bmi()` function for use in passive imputation to calculate `bmi`. The predictor matrix is set up so that only `age`, `sex`, `hr` and `wr` are permitted to impute `hm` and `wm`.

Figure 7.6 is a diagnostic plot to check whether the imputations maintain the relation between the measured and the self-reported data. The plot is identical to Figure 7.4, except that the imputed data from the survey data (in red) have been added. Imputations have been taken from the first imputed dataset. The figure shows that the red and blue dots are similar in terms of

Table 7.7: Obesity prevalence estimate (%) and standard error (se) in the survey data ($n = 803$), reported (observed data) and corrected (imputed).

			Reported		Corrected	
Sex	Age	n	%	se	%	se
Male	18–29	69	8.7	3.4	9.4	3.9
	30–39	73	11.0	3.7	15.7	5.0
	40–49	66	9.1	3.6	12.5	4.8
	50–59	91	20.9	4.3	25.4	5.2
	60–75	101	7.9	2.7	15.6	4.2
	18–75	400	11.7	1.6	16.0	2.0
Female	18–29	68	14.7	4.3	16.3	5.7
	30–39	69	26.1	5.3	28.4	6.6
	40–49	68	19.1	4.8	25.4	6.1
	50–59	81	25.9	4.9	32.8	6.0
	60–75	117	11.1	2.9	17.1	4.6
	18–75	403	18.6	1.9	23.0	2.4
M & F	18–75	803	15.2	1.3	19.5	1.5

location and spread. Observe that BMI in the survey data is slightly higher. The very small difference between the smoothed lines across all measured BMI values confirms this notion. We conclude that the relation between self-reported and measured BMI as observed in the calibration data successfully "migrated" to the survey data.

Table 7.7 contains the prevalence estimates based on the survey data given for self-report and corrected for self-reporting bias. The estimates themselves are variable and have large standard errors. It is easy to infer that the size of the correction depends on age. Note that the standard errors of the corrected estimates are always larger than for the self-report. This reflects the information lost due to the correction. To obtain an equally precise estimate, the sample size of the study with only self-reports needs to be larger than the sample size of the study with direct measures.

7.3.6 Conclusion

Predictive equations to correct for self-reporting bias will only work if the percentage of explained variance is very high. In the general case, they have a systematic downward bias, which makes them unsuitable as correction methods. The remedy is to explicitly account for the residual distribution. We have done so by applying multiple imputation to impute measured height and weight. In addition, multiple imputation produces the correct standard errors of the prevalence estimates.

7.4 Enhancing comparability

7.4.1 Description of the problem

Comparability of data is a key problem in international comparisons and meta analysis. The problem of comparability has many sides. An overview of the issues and methodologies can be found in Van Deth (1998), Harkness et al. (2002), Salomon et al. (2004), King et al. (2004), Matsumoto and Van de Vijver (2010) and Chevalier and Fielding (2011).

This section addresses just one aspect, incomparability of the data obtained on survey items with different questions or response categories. This is a very common problem that hampers many comparisons.

One of the tasks of the European Commission is to provide insight into the level of disability of the populations in each of the 27 member states of the European Union. Many member states conduct health surveys, but the precise way in which disability is measured are very different. For example, The U.K. Health Survey contains a question *How far can you walk without stopping/experiencing severe discomfort, on your own, with aid if normally used?* with response categories "can't walk," "a few steps only," "more than a few steps but less than 200 yards" and "200 yards or more." The Dutch Health Interview Survey contains the question *Can you walk 400 metres without resting (with walking stick if necessary)?* with response categories "yes, no difficulty," "yes, with minor difficulty," "yes, with major difficulty" and "no." Both items obviously intend to measure the ability to walk, but it is far from clear how an answer on the U.K. item can be compared with one on the Dutch item.

Response conversion (Van Buuren et al., 2005) is a way to solve this problem. The technique transforms responses obtained on different questions onto a common scale. Where this can be done, comparisons can be made using the common scale. The actual data transformation can be repeatedly done on a routine basis as new information arrives. The construction of *conversion keys* is only possible if enough overlapping information can be identified. Keys have been constructed for dressing disability (Van Buuren et al., 2003), personal care disability, sensory functioning and communication, physical well-being (Van Buuren and Tennant, 2004), walking disability (Van Buuren et al., 2005) and physical activity (Hopman-Rock et al., 2012).

This section presents an extension based on multiple imputation. The approach is more flexible and more general than response conversion. Multiple imputation does not require a common unidimensional latent scale, thereby increasing the range of applications.

7.4.2 Full dependence: Simple equating

In principle, the comparability problem is easy to solve if all sources would collect the same data. In practice, setting up and maintaining a centralized, harmonized data collection is easier said than done. Moreover, even where such efforts are successful, comparability is certainly not guaranteed (Harkness et al., 2002). Many factors contribute to the incomparability of data, but we will not go into details here.

In the remainder, we take an example of two bureaus that each collect health data on its own population. The bureaus use survey items that are similar, but not the same. The survey used by bureau A contains an item for measuring walking disability (item A):

> Are you able to walk outdoors on flat ground?
>
> 0: Without any difficulty
>
> 1: With some difficulty
>
> 2: With much difficulty
>
> 3: Unable to do

The frequencies observed in sample A are 242, 43, 15 and 0. There are six missing values. Bureau A produces a yearly report containing an estimate of the mean of the distribution of population A on item A. Assuming MCAR, a simple random sample and equal inter-category distances, we find $\hat{\theta}_{\mathrm{AA}} = (242 * 0 + 43 * 1 + 15 * 2)/300 = 0.243$, the disability estimate for population A using the method of bureau A.

The survey of bureau B contains item B:

> Can you, fully independently, walk outdoors (if necessary, with a cane)?
>
> 0: Yes, no difficulty
>
> 1: Yes, with some difficulty
>
> 2: Yes, with much difficulty
>
> 3: No, only with help from others

The frequencies observed in sample B are 145, 110, 29 and 8. There were no missing values reported by bureau B. Bureau B publishes the proportion of cases in category 0 as a yearly health measure. Assuming a simple random sample, $P(Y_{\mathrm{B}} = 0)$ is estimated by $\hat{\theta}_{\mathrm{BB}} = 145/292 = 0.497$, the health estimate for population B using the method of bureau B.

Note that $\hat{\theta}_{\mathrm{AA}}$ and $\hat{\theta}_{\mathrm{BB}}$ are different statistics calculated on different samples, and hence cannot be compared. On the surface, the problem is trivial and can be solved by just equating the four categories. After that is done, and we can apply the methods of bureau A or B, and compare the results. Such recoding to "make data comparable" is widely practiced.

Let us calculate the result using simple equating. To estimate walking disability in population B using the method of bureau A we obtain $\hat{\theta}_{\mathrm{BA}} = (145*0+110*1+29*2+8*3)/292 = 0.658$. Remember that the mean disability estimate for population A was equal to 0.243, so population B appears to have substantially more walking disability. The difference equals $\hat{\theta}_{\mathrm{BA}} - \hat{\theta}_{\mathrm{AA}} = 0.658 - 0.243 = 0.414$ on a scale from 0 to 3.

Likewise, we may estimate bureau's B health measure θ_{AB} in population A as $\hat{\theta}_{\mathrm{AB}} = 242/300 = 0.807$. Thus, over 80% of population A scores in category 0. This is substantially more than in population B, which was $\hat{\theta}_{\mathrm{BB}} = 145/292 = 0.497$.

So by equating categories both bureaus conclude that the healthier population is A, and by a fairly large margin. As we will see, this result is however highly dependent on assumptions that may not be realistic for these data.

7.4.3 Independence: Imputation without a bridge study

Let Y_{A} be the item of bureau A, and let Y_{B} be the item of bureau B. The comparability problem can be seen as a missing data problem, where Y_{A} is missing for population B, and where Y_{B} is missing for population A. This formulation suggest that we can use imputation to solve the problem, and calculate $\hat{\theta}_{\mathrm{AB}}$ and $\hat{\theta}_{\mathrm{BA}}$ from the imputed data.

Let's see what happens if we put `mice()` to work to solve the problem. We first create the dataset:

```
> fA <- c(242, 43, 15, 0, 6)
> fB <- c(145, 110, 29, 8)
> YA <- rep(ordered(c(0:3, NA)), fA)
> YB <- rep(ordered(c(0:3)), fB)
> Y <- rbind(data.frame(YA, YB = ordered(NA)), data.frame(YB,
      YA = ordered(NA)))
```

The data `Y` is a data frame with 604 rows and 2 columns: `YA` and `YB`. The missing data pattern is

```
> md.pattern(Y)
     YA  YB
292   0   1   1
300   1   0   1
  6   0   0   2
    298 306 604
```

The missing data pattern is unconnected (cf. Section 4.1.1), with no observations linking `YA` to `YB`. There are six records that contain no data at all.

For this problem, we monitor the behavior of a rank-order correlation, Kendall's τ, between `YA` and `YB`. This is not a standard facility in `mice()`, but

we can easily write a small function `micemill()` that calculates Kendall's τ after each iteration as follows.

```
> micemill <- function(n) {
      for (i in 1:n) {
          imp <<- mice.mids(imp)
          cors <- with(imp, cor(as.numeric(YA),
              as.numeric(YB), method = "kendall"))
          tau <<- rbind(tau, ra(cors, s = TRUE))
      }
  }
```

This function calls `mice.mids()` to perform just one iteration, calculates Kendall's τ, and stores the result. Note that the function contains two double assignment operators. This allows the function to overwrite the current `imp` and `tau` object in the global environment. This is a dangerous operation, and not really an example of good programming in general. However, we may now write

```
> tau <- NULL
> imp <- mice(Y, max = 0, m = 10, seed = 32662)
> micemill(10)
```

This code executes 10 iterations of the MICE algorithm. Normally, 10 iterations are enough, but—as we will see—that is not the case here. We can ask for 40 additional iterations by typing `micemill(40)` at the prompt. After any number of iterations, we may plot the trace lines of the MICE algorithm by

```
> plotit <- function() matplot(x = 1:nrow(tau),
      y = tau, ylab = expression(paste("Kendall's ",
          tau)), xlab = "Iteration", type = "l",
      lwd = 1, lty = 1:10, col = "black")
> plotit()
```

Figure 7.7 contains the trace plot of 50 iterations. The traces start near zero, but then freely wander off over a substantial range of the correlation. In principle, the traces could hit values close to $+1$ or -1, but that is an extremely unlikely event. The MICE algorithm obviously does not know where to go, and wanders pointlessly through parameter space. The reason that this occurs is that the data contain no information about the relation between Y_{A} and Y_{B}.

Despite the absence of any information about the relation between Y_{A} and Y_{B}, we can calculate $\hat{\theta}_{\text{AB}}$ and $\hat{\theta}_{\text{BA}}$ without a problem from the imputed data. We find $\hat{\theta}_{\text{AB}} = 0.500$ (SD: 0.031), which is very close to $\hat{\theta}_{\text{BB}}$ (0.497), and far from the estimate under simple equating (0.807). Likewise, we find $\hat{\theta}_{\text{BA}} = 0.253$ (SD: 0.034), very close to $\hat{\theta}_{\text{AA}}$ (0.243) and far from the estimate under

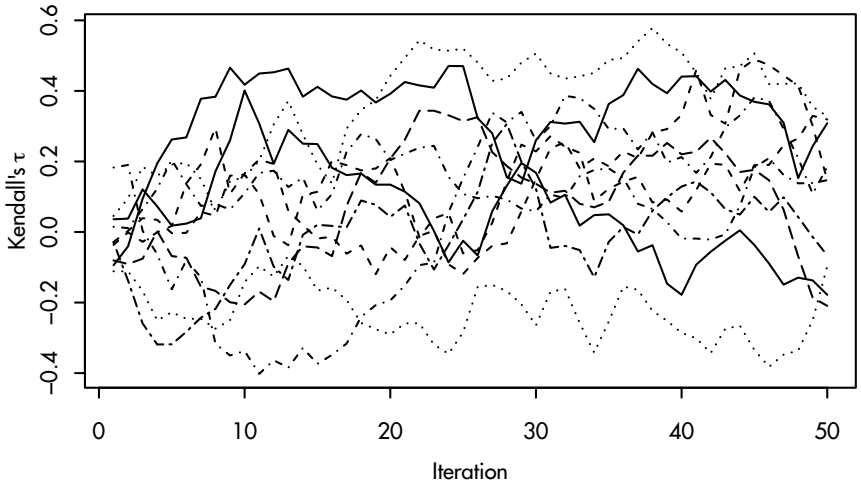


Figure 7.7: The trace plot of Kendall's τ for Y_{A} and Y_{B} using $m = 10$ multiple imputations and 50 iterations. The data contain no cases that have observations on both Y_{A} and Y_{B}.

equating (0.658). Thus, if we perform the analysis without any information that links the items, we consistently find no difference between the estimates for populations A and B, despite the huge variation in Kendall's τ.

We have now two estimates of $\hat{\theta}_{\text{AB}}$ and $\hat{\theta}_{\text{BA}}$. In particular, in Section 7.4.2 we calculated $\hat{\theta}_{\text{BA}} = 0.658$ and $\hat{\theta}_{\text{AB}} = 0.807$, whereas in the present section the results are $\hat{\theta}_{\text{BA}} = 0.253$ and $\hat{\theta}_{\text{AB}} = 0.500$, respectively. Thus, both health measures are very dissimilar due to the assumptions made. The question is which method yields results that are closer to the truth.

7.4.4 Fully dependent or independent?

Equating categories is equivalent to assuming that the pairs are 100% concordant. In that case Kendall's τ is equal to 1. Figure 7.7 illustrates that it is extremely unlikely that $\tau = 1$ will happen by chance. On the other hand, the two items look very similar, so Kendall's τ could be high on that basis. In order to make progress, we need to look at the data, and estimate τ.

Suppose that item Y_{A} and Y_{A} had both been administered to an external sample, called sample E. Table 7.8 contains the contingency table of Y_{A} and Y_{B} in sample E, taken from Van Buuren et al. (2005). Although there is a strong relation between Y_{A} and Y_{B}, the contingency table is far from diagonal. For example, category 1 of Y_{B} has 110 observations, whereas category 1 of Y_{A} contains only 68 persons. The table is also not symmetric, and suggests that Y_{A} is more difficult than Y_{B}. In other words, a given score on Y_{A} corresponds to more walking disability compare to the same score on Y_{B}. Kendall's

Table 7.8: Contingency table of responses on Y_A and Y_B in an external sample E ($n = 292$).

			Y_B		
Y_A	0	1	2	3	Total
0	128	45	3	2	178
1	13	45	10	0	68
2	3	20	14	5	42
3	0	0	1	1	2
NA	1	0	1	0	2
Total	145	110	29	8	292

τ is equal to 0.57, so about 57% of the pairs are concordant. This is far better than chance (0%), but also far worse than 100% concordance implied by simple equating. Thus even though the four response categories of Y_A and Y_B look similar, the information from sample E suggests that there are large and systematic differences in the way the items work. Given these data, the assumption of equal categories is in fact untenable. Likewise, the solution that assumes independence is also unlikely.

The implication is that both estimates of θ_{AB} and θ_{BA} presented thus far are doubtful. At this stage, we cannot yet tell which of the estimates is the better one.

7.4.5 Imputation using a bridge study

We will now rerun the imputation, but with sample E appended to the data from the sample for populations A and B. Sample E acts as a bridge study that connects the missing data patterns from samples A and B. The combined data are available in `mice` as the dataset `walking`. The missing data pattern is

```
> md.pattern(walking)
    sex age src  YA  YB
290   1   1   1   1   1   0
294   1   1   1   0   1   1
300   1   1   1   1   0   1
  6   1   1   1   0   0   2
      0   0   0 300 306 606
```

Observe that `YA` and `YB` are now connected by 290 records from the bridge study on sample E. We assume that the data are missing at random. More specifically, the conditional distributions of Y_A and Y_B given the other item is equivalent across the three sources. Let S be an administrative variable taking on values A, B and E for the three sources. The assumptions are

$$P(Y_A|Y_B, X, S = B) \quad = \quad P(Y_A|Y_B, X, S = E) \tag{7.1}$$

$$P(Y_{\mathrm{B}}|Y_{\mathrm{A}}, X, S = A) \quad = \quad P(Y_{\mathrm{B}}|Y_{\mathrm{A}}, X, S = E) \tag{7.2}$$

where X contains any relevant covariates, like age and sex, and/or interaction terms. In other words, the way in which Y_{A} depends on Y_{B} and X is the same in sources B and E. Likewise, the way in which Y_{B} depends on Y_{A} and X is the same in sources A and E. The inclusion of such covariates allows for various forms of differential item functioning (Holland and Wainer, 1993).

The two assumptions need critical evaluation. For example, if the respondents in source $S = E$ answered the items in a different language than the respondents in sources A or B, then the assumption may not be sensible unless one has great faith in the translation. It is perhaps better then to search for a bridge study that is more comparable.

Note that it is only required that the conditional distributions are identical. The imputations remain valid when the samples have different marginal distributions. For efficiency reasons and stability, it is generally advisable to have match samples with similar distribution, but it is not a requirement. The design is known as the *common-item nonequivalent groups* design (Kolen and Brennan, 1995) or the *non-equivalent group anchor test (NEAT)* design (Dorans, 2007).

Multiple imputation on the dataset `walking` is straightforward.

```
> tau <- NULL
> imp <- mice(walking, max = 0, m = 10, seed = 92786)
> pred <- imp$pred
> pred[, c("src", "age", "sex")] <- 0
> imp <- mice(walking, max = 0, m = 10, seed = 92786,
      pred = pred)
> micemill(20)
> plotit()
```

The behavior of the trace plot is very different now (cf. Figure 7.8). After the first few iterations, the trace lines consistently move around a value of approximately 0.53, with a fairly small range. Thus, after five iterations, the conditional distributions defined by sample E have percolated into the imputations for item A (in sample B) and item B (in sample A).

The behavior of the samplers is dependent on the relative size of the bridge study. In these data, the bridge study is about one third of the total data. If the bridge study is small relative to the other two data sources, the sampler may be slow to converge. As a rule of the thumb, the bridge study should be at least 10% of the total sample size. Also, carefully monitor convergence of the most critical linkages using association measures.

Note that we can also monitor the behavior of $\hat{\theta}_{\mathrm{AB}}$ and $\hat{\theta}_{\mathrm{BA}}$. In order to calculate $\hat{\theta}_{\mathrm{AB}}$ after each iteration we add two statements to the `micemill()` function:

```
> props <- with(imp, mean(YB[src == "A"] == "0"))
> thetaAB <<- rbind(thetaAB, ra(props, s = TRUE))
```

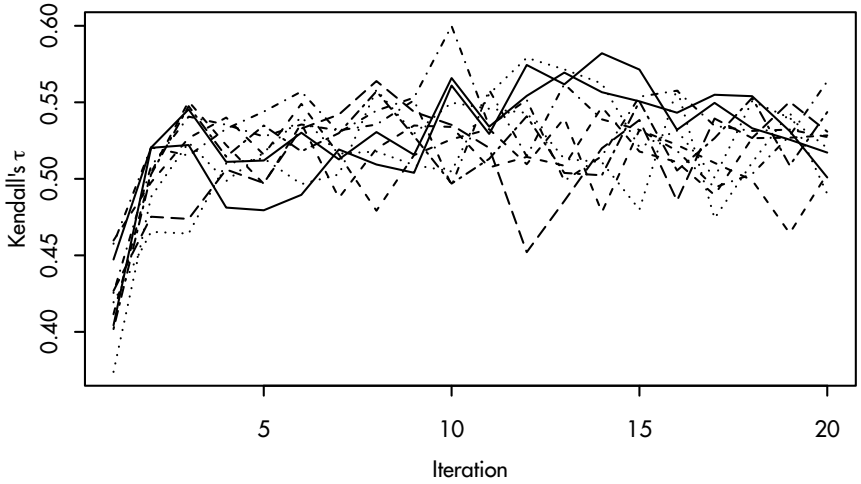


Figure 7.8: The trace plot of Kendall's τ for Y_{A} and Y_{B} using $m = 10$ multiple imputations and 20 iterations. The data are linked by the bridge study.

The results are assembled in the variable `thetaAB` in the working directory. This variable should be initialized as `thetaAB <- NULL` before milling.

It is possible that the relation between Y_{A} and Y_{B} depends on covariates, like age and sex. If so, including covariates into the imputation model allows for differential item functioning across the covariates. It is perfectly possible to change the imputation model between iterations. For example, after the first 20 iterations (where we impute Y_{A} from Y_{B} and vice versa) we add age and sex as covariates, and do another 20 iterations. This goes as follows:

```
> tau <- NULL
> thetaAB <- NULL
> imp <- mice(walking, max = 0, m = 10, seed = 99786)
> oldpred <- pred <- imp$pred
> pred[, c("src", "age", "sex")] <- 0
> imp <- mice(walking, max = 0, m = 10, seed = 99786,
      pred = pred)
> micemill(20)
> pred <- oldpred
> pred[, c("src")] <- 0
> imp <- mice(walking, max = 0, m = 10, pred = pred)
> micemill(20)
```

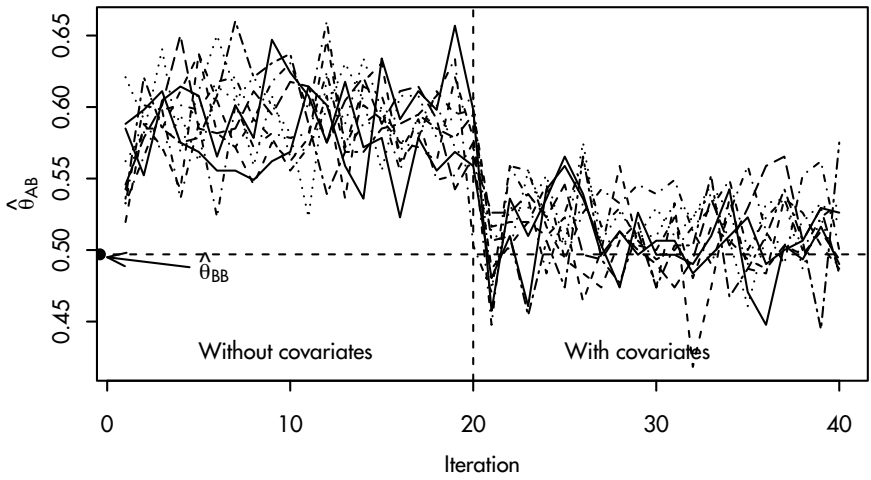


Figure 7.9: Trace plot of $\hat{\theta}_{\text{AB}}$ (proportion of sample A that scores in category 0 of item B) after multiple imputation ($m = 10$), without covariates (iteration 1–20), and with covariates age and sex as part of the imputation model (iterations 21–40).

7.4.6 Interpretation

Figure 7.9 plots the traces of MICE algorithm, where we calculated θ_{AB}, the proportion of sample A in category 0 of item B. Without covariates, the proportion is approximately 0.58. Under equating, this proportion was found to be equal to 0.807 (cf. Section 7.4.2). The difference between the old (0.807) and the new (0.580) estimate is dramatic. After adding age and sex to the imputation model, θ_{AB} drops further to about 0.510, close to θ_{BB}, the estimate for population B (0.497).

Table 7.9 summarizes the estimates from the four analyses. Large differences are found between population A and B when we simply assume that the four categories of both items are identical (simple equating). In this case, population A appears much healthier by both measures. In constrast, if we assume independence between Y_{A} and Y_{B}, all differences vanish, so now it appears that the populations A and B are equally healthy. The solutions based on multiple imputation strike a balance between these extremes. Population A is considerably healthier than B on the item mean statistic (0.243 versus 0.451). However, the difference is much smaller on the proportion in category 0, especially after taking age and sex into account. The solutions based on multiple imputation are preferable over the first two because they have taken the relation between items A and B into account.

Which of the four estimates is "best?" The method of choice is multiple imputation including the covariates. This method not only accounts for the

Table 7.9: Disability and health estimates for populations A and B under four assumptions. $\hat{\theta}_{\mathrm{AA}}$ and $\hat{\theta}_{\mathrm{BA}}$ are the item means on item A for samples A and B, respectively. $\hat{\theta}_{\mathrm{AB}}$ and $\hat{\theta}_{\mathrm{BB}}$ are the proportions of cases into category 0 of item B for samples A and B, respectively. MI-multiple imputation.

Assumption	$\hat{\theta}_{\mathrm{AA}}$	$\hat{\theta}_{\mathrm{BA}}$	$\hat{\theta}_{\mathrm{AB}}$	$\hat{\theta}_{\mathrm{BB}}$
Simple equating	0.243	0.658	0.807	0.497
Independence	0.243	0.253	0.500	0.497
MI (no covariate)	0.243	0.450	0.580	0.497
MI (covariate)	0.243	0.451	0.510	0.497

relation between Y_A and Y_B, but also incorporates the effects of age and sex. Consequently, the method provides estimates with the lowest bias in θ_{AB} and θ_{BA}.

7.4.7 Conclusion

Incomparability of data is a key problem in many fields. It is natural for scientists to adapt, refine and tweak measurement procedures in the hope of obtaining better data. Frequent changes, however, will hamper comparisons.

Equating categories is widely practiced to "make the data comparable." It is often not realized that recoding and equating data amplify differences. The degree of exaggeration is inversely related to Kendall's τ. For the item mean statistic, the difference in mean walking disability after equating is about twice the size of that under multiple imputation. Also, the estimate of 0.807 after simple equating is a gross overestimate. Overstated differences between populations may spur inappropriate interventions, sometimes with substantial financial consequences. Unless backed up by appropriate data, equating categories is not a solution.

The section used multiple imputation as a natural and attractive alternative. The first major application of multiple imputation addressed issues of comparability (Clogg et al., 1991). The advantage is that bureau A can interpret the information of bureau B using the scale of bureau A, and vice versa. The method provides possible contingency tables of items A and B that could have been observed if both had been measured.

Dorans (2007) describes techniques for creating valid equating tables. Such tables convert the score of instrument A into that of instrument B, and vice versa. The requirements for constructing such tables are extremely high: the measured constructs should be equal, the reliability should be equal, the conversion of B to A should be the inverse of that from B to A (symmetry), it should not matter whether A or B is measured and the table should be independent of the population. Holland (2007) presents a logical sequence of linking methods that progressively moves toward higher forms of equating. Multiple imputation in general fails on the symmetry requirement, as it pro-

duces m scores on B for one score of A, and thus cannot be invertible. The method as presented here can be seen as a first step toward obtaining formal equating of test items. It can be improved by correcting for the reliabilities of both items. This is an area of future research.

For simplicity, the statistical analyses used only one bridge item. In general, better strategies are possible. It is wise to include as many bridge items as there are. Also, linking and equating at the sub-scale and scale levels could be done (Dorans, 2007). The double-coded data could also comprise a series of vignettes (Salomon et al., 2004). The use of such strategies in combination with multiple imputation has yet to be explored.

7.5 Exercises

1. *Contingency table.* Adapt the `micemill()` function for the `walking` data so that it prints out the contingency table of Y_{A} and Y_{B} of the first imputation at each iteration. How many statements do you need?

2. *Pool τ.* Find out what the variance of Kendall's τ is, and construct its 95% confidence intervals under multiple imputation. Use the auxiliary function `pool.scalar()` for pooling.

3. *Covariates.* Calculate the correlation between age and the items A and B under two imputation models: one without covariates, and one with covariates. Which of the correlations is higher? Which solution do you prefer? Why?

4. *Heterogeneity.* Kendall's τ in the source E is 0.57 (cf. Section 7.4.4). The average of the sampler is slightly lower (Figure 7.8). Adapt the `micemill()` function to calculate the τ-values separately for the three sources. Which population has the lowest τ-values?

5. *Sample size.* Repeat the previous exercise, but with the samples for A and B taken 10 times as large. Does the sample size have an effect on convergence? If so, can you come up with an explanation? (Hint: Think of how τ is calculated.)

6. *True values.* For sample B, we do actually have the data on Item A from sample E. Calculate the "true" value θ_{BA}, and compare it with the simulated values. How do these values compare? Should these values be the same? If they are different, what could be the explanations? How could you reorganize the `walking` data so that no iteration is needed?

Chapter 8

Selection issues

Chapter 7 concentrated on problems with the columns of the data matrix. Chapter 8 changes the perspective to the rows. An important consequence of nonresponse and missing data is that the remaining sample may not be representative anymore. Multiple imputation of entire blocks of variables (Section 8.1) can be useful to adjust for selective loss of cases in panel and cohort studies. Section 8.2 takes this idea a step further by appended and imputing new synthetic records to the data. This can also work for cross-sectional studies.

8.1 Correcting for selective drop-out

Panel attrition is a problem that plagues all studies in which the same people are followed over time. People who leave the study are called *drop-outs*. The persons who drop out may be systematically different from those who remain, thus providing an opportunity for bias. This section assumes that the drop-out mechanism is MAR and that the parameters of the complete-data model and the response mechanism are distinct (cf. Section 2.2.5). Techniques for nonignorable drop-outs are described by Little (1995), Diggle et al. (2002), Daniels and Hogan (2008) and Wu (2010).

8.1.1 POPS study: 19 years follow-up

The Project on Preterm and Small for Gestational Age Infants (POPS) is an ongoing collaborative study in the Netherlands on the long-term effect of prematurity and dysmaturity on medical, psychological and social outcomes. The cohort was started in 1983 and enrolled 1338 infants with a gestational age below 32 weeks or with a birth weight of below 1500 grams (Verloove-Vanhorick et al., 1986). Of this cohort, 312 infants died in the first 28 days, and another 67 children died between the ages of 28 days and 19 years, leaving 959 survivors at the age of 19 years. Intermediate outcome measures from earlier follow-ups were available for 89% of the survivors at age 14 ($n = 854$), 77% at age 10 ($n = 712$), 84% at age 9($n = 813$), 96% at age 5 ($n = 927$) and 97% at age 2($n = 946$).

Table 8.1: Count (percentage) of various factors for three response groups. Source: Hille et al. (2005).

	All responders		Full responders		Postal responders		Non-responders	
n	959		596		109		254	
Sex								
Boy	497	(51.8)	269	(45.1)	60	(55.0)	168	(66.1)
Girl	462	(48.2)	327	(54.9)	49	(45.0)	86	(33.9)
Origin								
Dutch	812	(84.7)	524	(87.9)	96	(88.1)	192	(75.6)
Non-Dutch	147	(15.3)	72	(12.1)	13	(11.9)	62	(24.4)
Maternal education								
Low	437	(49.9)	247	(43.0)	55	(52.9)	135	(68.2)
Medium	299	(34.1)	221	(38.5)	31	(29.8)	47	(23.7)
High	140	(16.0)	106	(18.5)	18	(17.3)	16	(8.1)
Social economic level								
Low	398	(42.2)	210	(35.5)	48	(44.4)	140	(58.8)
Medium	290	(30.9)	193	(32.6)	31	(28.7)	66	(27.7)
High	250	(26.7)	189	(31.9)	29	(26.9)	32	(13.4)
Handicap status at age 14 years								
Normal	480	(50.8)	308	(51.7)	42	(38.5)	130	(54.2)
Impairment	247	(26.1)	166	(27.9)	36	(33.0)	45	(18.8)
Mild	153	(16.2)	101	(16.9)	16	(14.7)	36	(15.0)
Severe	65	(6.9)	21	(3.5)	15	(13.8)	29	(12.1)

To study the effect of drop-out, Hille et al. (2005) divided the 959 survivors into three response groups:

1. *Full responders* were examined at an outpatient clinic and completed the questionnaires ($n = 596$);
2. *Postal responders* only completed the mailed questionnaires ($n = 109$);
3. *Non-responders* did not respond to any of the mailed requests or telephone calls, or could not be traced ($n = 254$).

8.1.2 Characterization of the drop-out

Of the 254 non-responders, 38 children (15%) did not comply because they were "physically or mentally unable to participate in the assessment." About half of the children (132, 52%) refused to participate. No reason for drop-out was known for 84 children (33%).

Table 8.1 lists some of the major differences between the three response groups. Compared to the postal and non-responders, the full response group consists of more girls, contains more Dutch children, has higher educational

and social economic levels and has fewer handicaps. Clearly, the responders form a highly selective subgroup in the total cohort.

Differential drop-out from the less healthy children leads to an obvious underestimate of disease prevalence. For example, the incidence of handicaps would be severely underestimated if based on data from the full responders only. In addition, selective drop-out could bias regression parameters in predictive models if the reason for drop-out is related to the outcome of interest. This may happen, for example, if we try to predict handicaps at the age of 19 years from the full responders only. Thus, statistical parameters may be difficult to interpret in the presence of selective drop-out.

8.1.3 Imputation model

The primary interest of the investigators focused on 14 different outcomes at 19 years: cognition, hearing, vision, neuromotor functioning, ADHD, respiratory symptoms, height, BMI, health status (Health Utilities Index Mark 3), perceived health (London Handicap Scale), coping, self-efficacy, educational attainment and occupational activities. Since it is inefficient to create a multiply imputed dataset for each outcome separately, the goal is to construct one set of imputed data that is used for all analyses.

For each outcome, the investigator created a list of potentially relevant predictors according to the predictor selection strategy set forth in Section 5.3.2. In total, this resulted in a set of 85 unique variables. Only 4 of these were completely observed for all 959 children. Moreover, the information provided by the investigators was coded (in Microsoft Excel) as a 85×85 predictor matrix that is used to define the imputation model.

Figure 8.1 shows a miniature version of the predictor matrix. The dark cell indicates that the column variable is used to impute the row variable. Note the four complete variables with rows containing only zeroes. There are three blocks of variables. The first nine variables (Set 1: `geslacht`–`sga`) are potential confounders that should be controlled for in all analyses. The second set of variables (Set 2: `grad.t`–`sch910r`) are variables measured at intermediate time points that appear in specific models. The third set of variables (Set 3: `iq`–`occrec`) are the incomplete outcomes of primary interest collected at the age of 19 years. The imputation model is defined such that:

1. All variables in Set 1 are used as predictors to impute Set 1, to preserve relations between them;

2. All variables in Set 1 are used as predictors to impute Set 3, because all variables in Set 1 appear in the complete-data models of Set 3;

3. All variables in Set 3 are used as predictors to impute Set 3, to preserve the relation between the variables measured at age 19;

4. Selected variables in Set 2 that appear in complete-data models are used as predictors to impute specific variables in Set 3;

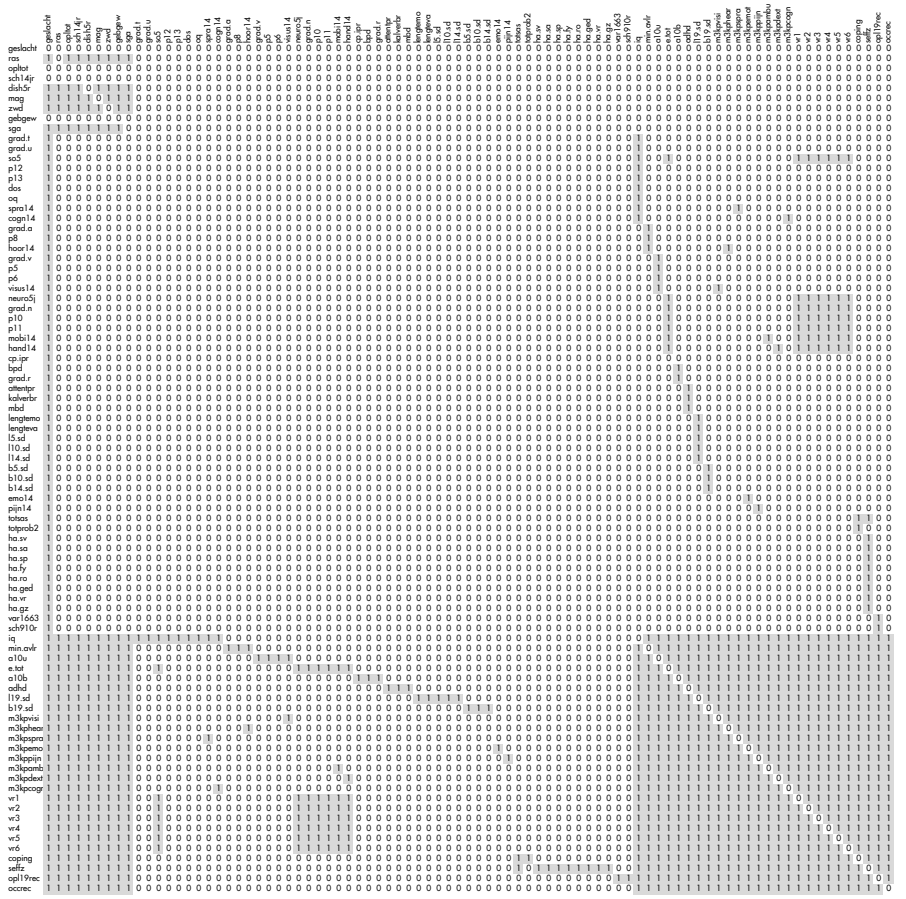


Figure 8.1: The 85×85 predictor matrix used in the POPS study. The gray parts signal the column variables that are used to impute the row variable.

5. Selected variables in Set 3 are "mirrored" to impute incomplete variables in Set 2, so as to maintain consistency between Set 2 and Set 3 variables;

6. The variable `geslacht` (sex) is included in all imputation models.

This setup of the predictor matrix avoids fitting unwieldy imputation models, while maintaining the relations of scientific interest.

8.1.4 A degenerate solution

The actual imputations can be produced by

```
> imp1 <- mice(data, pred = pred, maxit = 20, seed = 51121)
```

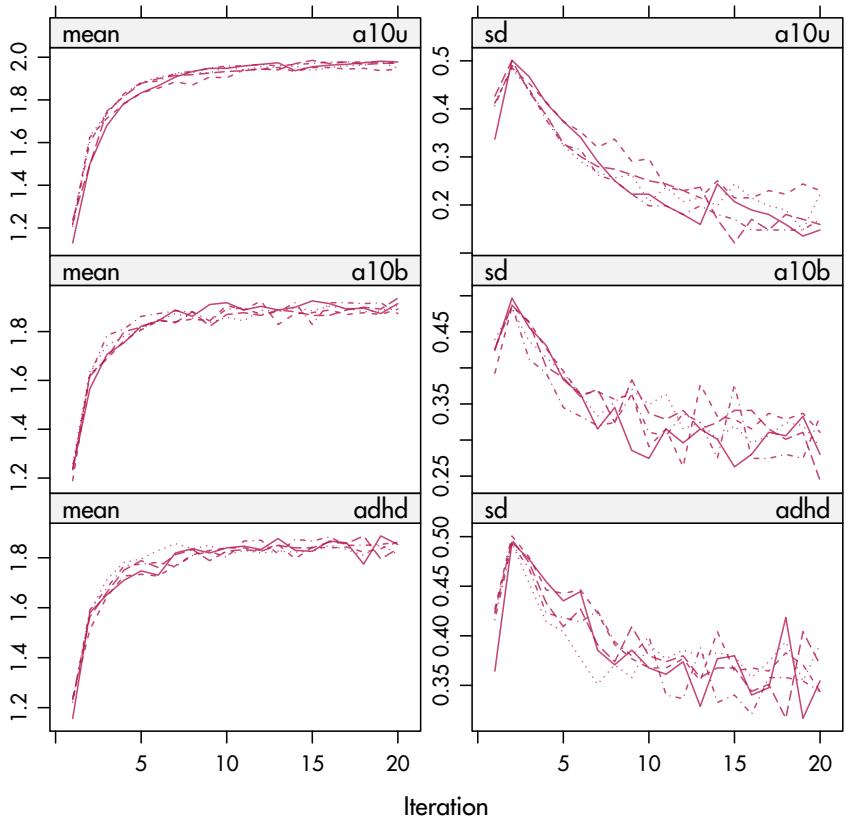


Figure 8.2: Trace lines of the MICE algorithm for three binary variables illustrating problematic convergence.

The number of iterations is set to 20 because the traces lines from MICE algorithm show strong initial trends and slow mixing. Figure 8.2 plots the trace lines of three binary variables `a10u` (visual disability), `a10b` (asthma, chronic bronchitis and CARA) and `adhd` (Attention deficit hyperactivity disorder). This figure is produced by

```
> plot(imp1, c("a10u", "a10b", "adhd"), col = mdc(5),
      lty = 1:5)
```

The behavior of these trace lines looks suspect. The lines connect the means (left side) of the synthetic values, and all converge to an asymptotic value around 1.9. Since the categories are coded as `1=no problem` and `2=problem`, a value of 1.9 actually implies that 90% of the nonresponders have a prob-

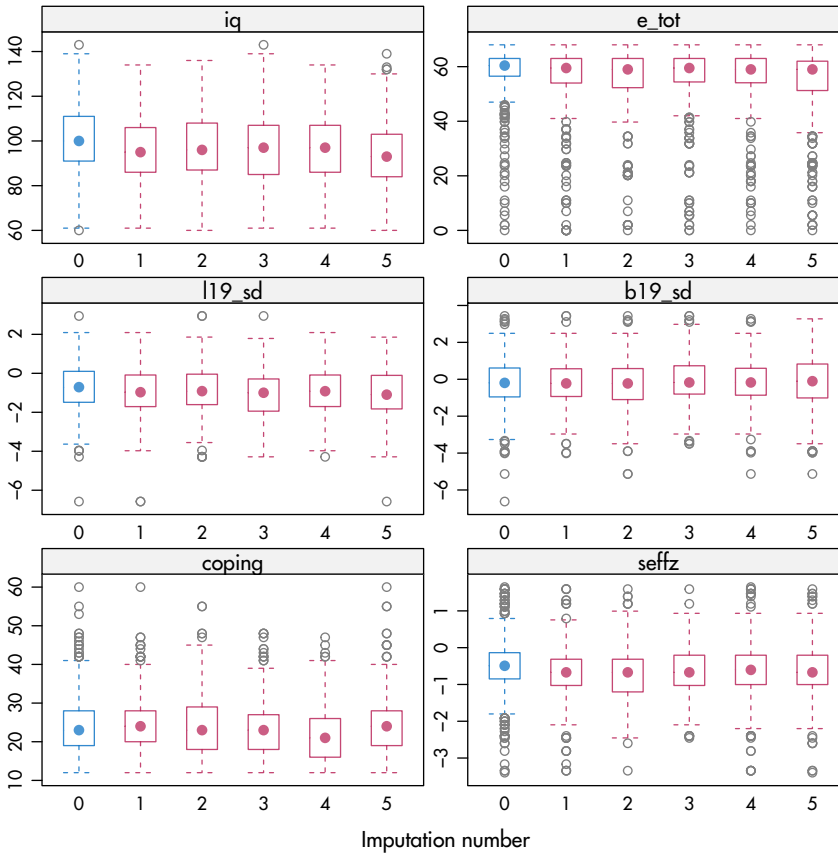


Figure 8.3: Distributions (observed and imputed values) of six outcome variables at 19 years in the POPS study.

lem. The prevalence in the full responders is equal to 1.5%, 8.0% and 4.7%, respectively. Although the nonresponse group is expected to have more health problems, 90% is clearly beyond any reasonable value. What we see here is a degenerate solution.

8.1.5 A better solution

The source of the problem is that all variables in Set 3 are used as predictors to impute Set 3. The MICE algorithm tends to move the imputations into a part of the parameter space where information is extremely sparse. A simple way to alleviate the problem is to remove the dependencies from Set 3. This can be done by adapting the predictor matrix as follows:

Table 8.2: Estimated percentage (95% CI) of three health problems at 19 years in the POPS study, uncorrected and corrected for selective drop-out.

	n_{obs}	Full responders		n	All children	
Severe visual handicap	690	1.4	(0.5–2.3)	959	4.7	(1.0–10.9)
Asthma, bronchitis, CARA	690	8.0	(5.9–10.0)	959	9.2	(6.9–11.2)
ADHD	666	4.7	(3.1–6.3)	959	5.7	(3.8–10.8)

```
> pred2 <- pred
> pred2[61:86, 61:86] <- 0
> imp2 <- mice(data, pred = pred2, maxit = 20, seed = 51121)
```

These statements produce imputations with marginal distributions much closer to the observed data. Also, the trace lines now show normal behavior (not shown). Convergence occurs rapidly in about 5–10 iterations.

Figure 8.3 is produced by

```
> bwplot(imp2, iq + e_tot + l19_sd + b19_sd + coping +
      seffz ~ .imp, layout = c(2, 3))
```

The figure displays the distributions of six numeric outcomes at 19 years. Both observed and imputed data are plotted by means of box-and-whisker plots. Observe that there are systematic differences for some outcomes. The nonrespondents have lower IQ scores (`iq`), are slightly shorter (`l19sd`) and have slightly lower self-efficacy scores (`seffz`).

8.1.6 Results

Table 8.2 provides estimates of the percentage of three health problems, both uncorrected and corrected for selective drop-out. As expected, all estimates are adjusted upward. Note that the prevalence of visual problems tripled to 4.7% after correction. While this increase is substantial, it is well within the range of odds ratios of 2.6 and 4.4 reported by Hille et al. (2005). The adjustment shows that prevalence estimates in the whole group can be substantially higher than in the group of full responders. Hille et al. (2007) provide additional and more detailed results.

8.1.7 Conclusion

Many studies are plagued by selective drop-out. Multiple imputation provides an intuitive way to adjust for drop-out, thus enabling estimation of statistics relative to the entire cohort rather than the subgroup. The method assumes MAR. The formulation of the imputation model requires some care. Section 8.1.3 outlines a simple strategy to specify the predictor matrix to fit an imputation model for multiple uses. This methodology is easily adapted to other studies.

Section 8.1.4 illustrates that multiple imputation is not without potential danger. The imputations produced by the initial model were far off. This underlines the importance of diagnostic evaluation of the imputed data. Section 8.1.5 described a solution for this problem. A disadvantage of this solution is that it preserves the relations between the variables in Set 3 only insofar as they are related through their common predictors. These relations may thus be attenuated. Alternatives to counter this problem (e.g., by specifying a triangular pattern in the predictor matrix) were not successful, as it appeared that the results were highly dependent on the sequence of the variables in the model. This is clearly an area that requires further investigation.

8.2 Correcting for nonresponse

This section describes how multiple imputation can be used to "make a sample representative." Weighting to known population totals is widely used to correct for nonresponse (Bethlehem, 2002; Särndal and Lundström, 2005). Imputation is an alternative to weighting. Imputation provides fine-grained control over the correction process. Provided that the imputation method is confidence proper, estimation of the correct standard errors can be done using Rubin's rules. Note however that this is not without controversy: Marker et al. (2002, p. 332) criticize multiple imputation as "difficult to apply," "to require massive amounts of computation," and question its performance for clustered datasets and unplanned analyses. Weighting and multiple imputation can also be combined, as was done in the NHANES III imputation project (Khare et al., 1993; Schafer et al., 1996).

This section demonstrates an application in the situation where the nonresponse is assumed to depend on known covariates, and where the distribution of covariates in the population is known. The sample is augmented by a set of artificial records, the outcomes in this set are multiply imputed and the whole set is analyzed. Though the application assumes random sampling, it should not be difficult to extend the basic ideas to more complex sampling designs.

8.2.1 Fifth Dutch Growth Study

The Fifth Dutch Growth Study is a cross-sectional nationwide study of height, weight and other anthropometric measurements among children 0–21 years living in the Netherlands (Schönbeck et al., 2011). The goal of the study is to provide updated growth charts that are representative for healthy children. The study is an update of similar studies performed in the Netherlands in 1955, 1965, 1980 and 1997. A strong secular trend in height has been observed over the last 150 years, making the Dutch population the tallest in the world (Fredriks et al., 2000a). The growth studies yield essential information

needed to calibrate the growth charts for monitoring childhood growth and development. One of the parameters of interest is *final height*, the mean height of the population when fully grown around the age of 20 years.

The survey took place between May 2008 and October 2009. The sample was stratified into five regions: North (Groningen, Friesland, Drenthe), East (Overijssel, Gelderland, Flevoland), West (Noord-Holland, Zuid-Holland, Utrecht), South (Zeeland, Noord-Brabant, Limburg) and the four major cities (Amsterdam, Rotterdam, The Hague, Utrecht City). The way in which the children were sampled depended on age. Up to 8 years of age, measurements were performed during regular periodical health examinations. Children older than 9 years were sampled from the population register, and received a personal invitation from the local health care provider.

The total population was stratified into three ethnic subpopulations. Here we consider only the subpopulation of Dutch descent. This group consists of all children whose biological parents are born in the Netherlands. Children with growth-related diseases were excluded. The planned sample size for the Dutch subpopulation was equal to 14782.

8.2.2 Nonresponse

During data collection, it quickly became evident that the response in children older than 15 years was extremely poor, and sometimes fell even below 20%. Though substantial nonresponse was caused by lack of perceived interest by the children, we could not rule out the possibility of selective nonresponse. For example, overweight children may have been less inclined to participate. The data collection method was changed in November 2008 so that all children with a school class were measured. Once a class was selected, nonresponse of the pupils was very generally small. In addition, children were measured by special teams at two high schools, two universities and a youth festival. The sample was supplemented with data from two studies from Amsterdam and Zwolle.

8.2.3 Comparison to known population totals

The realized sample size was $n = 10030$ children aged 0–21 years (4829 boys, 5201 girls). The nonresponse and the changes in the design may have biased the sample. If the sample is to be representative for the Netherlands, then the distribution of measured covariates like age, sex, region or educational level should conform to known population totals. Such population totals are based on administrative sources and are available in STATLINE, the online publication system of Statistics Netherlands.

Table 8.3 compares the proportion of children within five geographical regions in the Netherlands per January 1, 2010, with the proportions in the sample. Geography is known to be related to height, with the 20-year-olds in the North being about 3 cm taller in the North (Fredriks et al., 2000a).

Table 8.3: Distribution of the population and the sample over five geographical regions by age. Numbers are column percentages. Source: Fifth Dutch Growth Study (Schönbeck et al., 2011).

Region	0–9 Years		10–13 Years		14–21 Years	
	Population	Sample	Population	Sample	Population	Sample
North	12	7	12	11	12	4
East	24	28	24	11	24	55
South	23	27	24	31	25	21
West	21	26	20	26	20	15
City	20	12	19	22	19	4

Table 8.4: Number of observed and imputed children in the sample by geographical regions and age. Source: Fifth Dutch Growth Study (Schönbeck et al., 2011).

Region	0–9 Years		10–13 Years		14–21 Years	
	n_{obs}	n_{imp}	n_{obs}	n_{imp}	n_{obs}	n_{imp}
North	389	400	200	75	143	200
East	1654	0	207	300	667	0
South	1591	0	573	0	767	0
West	1530	0	476	0	572	0
City	696	600	401	0	164	400
Total	5860	1000	1857	375	2313	600

There are three age groups. In the youngest children, the population and sample proportions are reasonably close in the East, South and West, but there are too few children from the North and the major cities. For children aged 10–13 years, there are too few children from the North and East. In the oldest children, the sample underrepresents the North and the major cities, and overrepresents the East.

8.2.4 Augmenting the sample

The idea is to augment the sample in such a way that it will be nationally representative, followed by multiple imputation of the outcomes of interest. Table 8.4 lists the number of the measured children. The table also reports the number of children needed to bring the sample close to the population distribution.

In total 1975 records are appended to the 10030 records of children who were measured. The appended data contain three complete covariates: region, sex and age in years. For example, for the combination (`North`, `0-9 years`) $n_{\text{imp}} = 400$ new records are created as follows. All 400 records have the region category `North`. The first 200 records are boys and the last 200 records are girls. Age is drawn uniformly from the range 0–9 years. The outcomes of

interest, like height and weight, are set to missing. Similar blocks of records are created for the other five categories of interest, resulting in a total of 1975 new records with complete covariates and missing outcomes.

The following R code creates a dataset of 1975 records, with four complete covariates (id, reg, sex, age) and four missing outcomes (hgt, wgt, hgt.z, wgt.z). The outcomes hgt.z and wgt.z are standard deviation scores (SDS), or Z-scores, derived from hgt and wgt, respectively, standardized for age and sex relative to the Dutch references (Fredriks et al., 2000a).

```
> nimp <- c(400, 600, 75, 300, 200, 400)
> regcat <- c("North", "City", "North", "East",
      "North", "City")
> reg <- rep(regcat, nimp)
> nimp2 <- floor(rep(nimp, each = 2)/2)
> nimp2[5:6] <- c(38, 37)
> sex <- rep(rep(c("boy", "girl"), 6), nimp2)
> minage <- rep(c(0, 0, 10, 10, 14, 14), nimp)
> maxage <- rep(c(10, 10, 14, 14, 21, 21), nimp)
> set.seed(42444)
> age <- runif(length(minage), minage, maxage)
> id <- 600001:601975
> pad <- data.frame(id, reg, age, sex, hgt = NA,
      wgt = NA, hgt.z = NA, wgt.z = NA)
> data2 <- rbind(data, pad)
```

8.2.5 Imputation model

Regional differences in height are not constant across age, and tend to be more pronounced in older children. Figure 8.4 displays mean height standard deviation scores by age and region. Children from the North are generally the tallest, while those from the South are shortest, but the difference varies somewhat with age. Children from the major cities are short at early ages, but relatively tall in the oldest age groups. Imputation should preserve these features in the data, so we need to include at least the age by region interaction into the imputation model. In addition, we incorporate the interaction between SDS and age, so that the relation between height and weight could differ across age. In R, we create interaction by the model.matrix() function as follows:

```
> na.opt <- options(na.action = na.pass)
> int <- model.matrix(~I(age - 10) * hgt.z + I(age -
      10) * wgt.z + age * reg, data = data2)[, -(1:9)]
> options(na.opt)
> data3 <- cbind(data2, int)
```

Since model.frame() uses listwise deletion, we need to temporarily change the option na.action=na.pass. The interactions involving missing data in

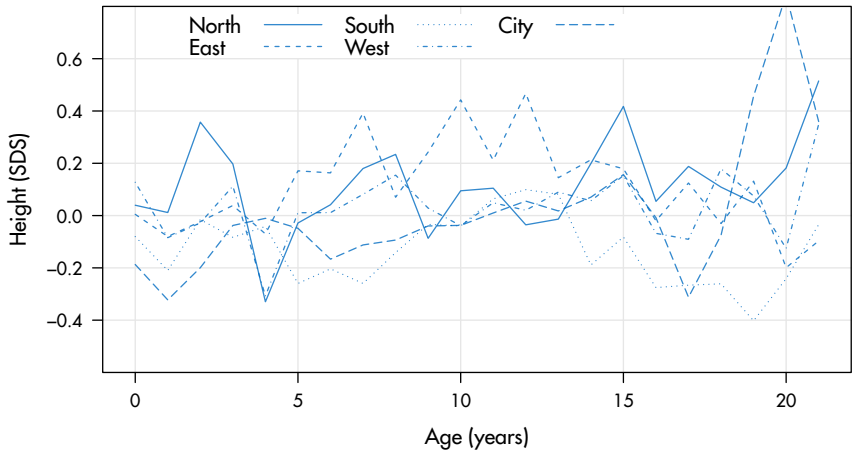


Figure 8.4: Height SDS by age and region of Dutch children. Source: Fifth Dutch Growth Study ($n = 10030$).

`hgt.z` or `wgt.z` will thus be set to `NA` rather than being removed. The selection operator `[,-(1:9)]` deletes the main effects. Height SDS and weight SDS are imputed by normal imputation. Absolute values in centimeters (cm) and kilograms (kg) are calculated after imputation from the imputed SDS. The setup of the imputation model uses passive imputation to update the interaction terms.

```
> ini <- mice(data3, maxit = 0)
> meth <- ini$meth
> meth["hgt"] <- ""
> meth["wgt"] <- ""
> meth["hgt.z"] <- "norm"
> meth["wgt.z"] <- "norm"
> meth["I(age - 10):hgt.z"] <- "~I(I(age-10)*hgt.z)"
> meth["I(age - 10):wgt.z"] <- "~I(I(age-10)*wgt.z)"
> pred <- ini$pred
> pred[, c("hgt", "wgt")] <- 0
> pred["hgt.z", c("id", "I(age - 10):hgt.z")] <- 0
> pred["wgt.z", c("id", "I(age - 10):wgt.z")] <- 0
> vis <- ini$vis[c(3, 5, 4, 6)]
> imp <- mice(data3, meth = meth, pred = pred, vis = vis,
      m = 10, maxit = 20, seed = 28107)
```

The SDS is approximately normally distributed with a mean of zero and a standard deviation of 1. We may thus use the linear normal model rather than

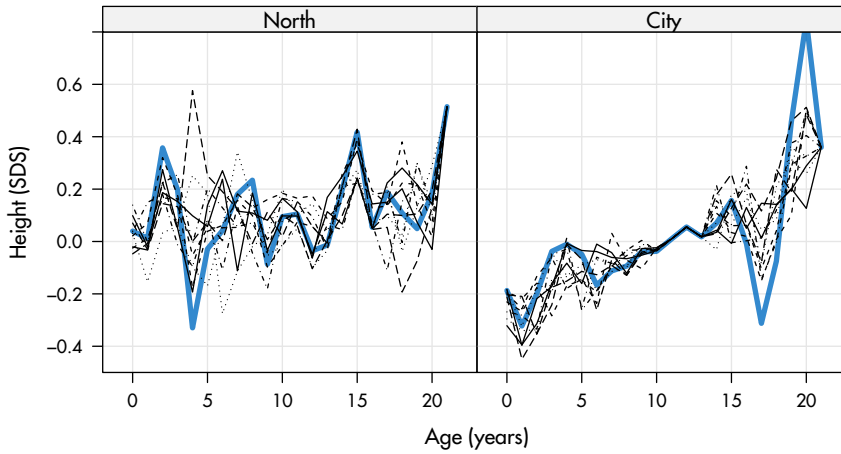


Figure 8.5: Mean height SDS per year for region `North` (left) and region `City` (right), for the original data ($n = 10030$) and 10 augmented datasets that correct for the nonresponse ($n = 12005$).

predictive mean matching. With this sample size, the linear model is much faster.

Figure 8.5 displays mean height SDS per year for regions `North` and `City` in the original and augmented data. The 10 imputed datasets show patterns in mean height SDS similar to those in the observed data. Because of the lower sample size, the means for region `North` are more variable than `City`. Observe also that the rising pattern in `City` is reproduced in the imputed data. No imputations were generated for the ages 10–13 years, which explains that the means of the imputed and observed data coincide. The imputations tend to smooth out sharp peaks at higher ages due to the low number of data points.

8.2.6 Influence of nonresponse on final height

Figure 8.6 displays the mean of fitted height distribution of the original and the 10 imputed datasets. Since children from the shorter population in the South are overrepresented, the estimates of final height from the sample (183.6 cm for boys, 170.6 cm for girls) are biased downward. The estimates calculated from the imputed data vary from 183.6 to 184.1 cm (boys) and 170.6 to 171.1 cm (girls). Thus, correcting for the nonresponse leads to final height estimates that are about 2 mm higher.

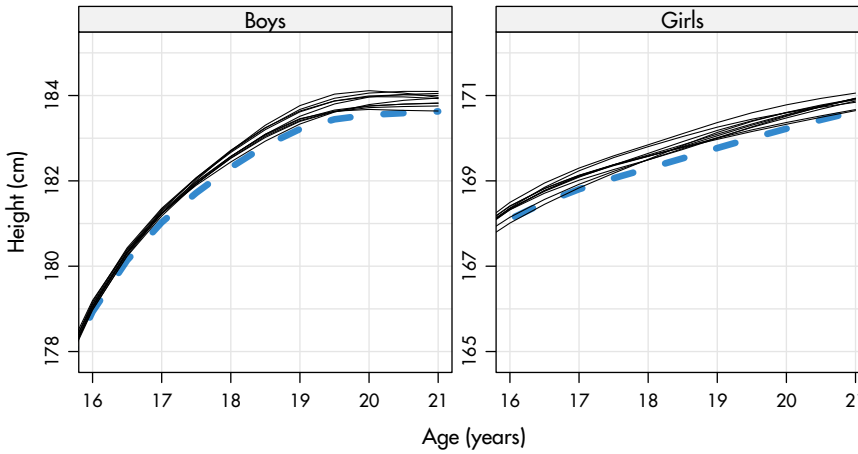


Figure 8.6: Final height estimates in Dutch boys and girls from the original sample ($n = 10030$) and 10 augmented samples ($n = 12005$) that correct for the nonresponse.

8.2.7 Discussion

The application as described here only imputes height and weight in Dutch children. It is straightforward to extend the method to impute additional outcomes, like waist or hip circumference.

The method can only correct for covariates whose distributions are known in both the sample and population. It does not work if nonresponse depends on factors for which we have no population distribution. However, if we have possession of nonresponse forms for a representative sample, we may use any covariates common to the responders and nonresponders to correct for the nonresponse using a similar methodology. The correction will be more successful if these covariates are related to the reasons for the nonresponse.

There are no accepted methods yet to calculate the number of extra records needed. Here we used 1975 new records to augment the existing 10030 records, about 16% of the total. This number of artificial records brought the covariate distribution in the augmented sample close to the population distribution without the need to discard any of the existing records. When the imbalance grows, we may need a higher percentage of augmentation. The estimates will then be based on a larger fraction of missing information, and may thus become unstable. Alternatively, we could sacrifice some of the existing records by taking a random subsample of strata that are overrepresented. It is not yet known whether this leads to more efficient estimates. More work is needed to validate the method and to compare it to traditional approaches based on weighting.

8.3 Exercises

1. *90th centile.* Repeat the analysis in Section 8.2.6 for final height. Study the effect of omitting the interaction effect from the imputation model. Are the effects on the 90th centile the same as for the mean?

2. *How many records?* Section 8.2.4 describes an application in which incomplete records are appended to create a representative sample. Develop a general strategy to determine the number of records needed to append.

This page intentionally left blank

Chapter 9

Longitudinal data

9.1 Long and wide format

Longitudinal data can be coded into "long" and "wide" formats. A wide dataset will have one record for each individual. The observations made at different time points are coded as different columns. In the wide format every measure that varies in time occupies a set of columns. In the long format there will be multiple records for each individual. Some variables that do not vary in time are identical in each record, whereas other variables vary across the records. The long format also needs a "time" variable that records the time in each record, and an "id" variable that groups the records from the same person.

A simple example of the wide format is

```
id age Y1 Y2
 1  14 28 22
 2  12 34 16
 3  ...
```

In the long format, this dataset looks like

```
id age  Y
 1  14 28
 1  14 22
 2  12 34
 2  12 16
 3  ...
```

Note that the concepts of long and wide are general, and also apply to cross-sectional data. For example, we have seen the long format before in Section 6.1.1, where it referred to stacked imputed data that was produced by the `complete()` function. The basic idea is the same.

Both formats have their advantages. If the data are collected on the same time points, the wide format has no redundancy or repetition. Elementary statistical computations like calculating means, change scores, age-to-age correlations between time points, or the t-test are easy to do in this format. The

long format is better at handling irregular and missed visits. Also, the long format has an explicit time variable available that can be used for analysis. Graphs and statistical analyses are easier in the long format.

Applied researchers often collect, store and analyze their data in the wide format. Classic ANOVA and MANOVA techniques for repeated measures and structural equation models for longitudinal data assume the wide format. Modern multilevel techniques and statistical graphs, however, work only from the long format. The distinction between the two formats is a first stumbling block for those new to longitudinal analysis.

Singer and Willett (2003) advise the data storing in both formats. The wide and the long formats can be converted into each other by a database operation. `R` and `Stata` have `reshape()` functions. In `SPSS` the wide-to-long conversion is done by the `VARSTOCASES` commands, and the long-to-wide conversion by `CASESTOVARS`. Both are available from the `Data Restructure...` menu. `SAS` uses `PROC TRANSPOSE` for this purpose.

Multiple imputation of longitudinal data is conveniently done when data are in the wide format. Apart from the fact that the columns are ordered in time, there is nothing special about the imputation problem. We may thus apply the techniques from the earlier chapters to longitudinal data. Section 9.2 discusses an imputation technique in the wide format in a clinical trial application with the goal of performing a statistical analysis according to the intention to treat (ITT) principle. The longitudinal character of the data helped specify the imputation model.

The wide-to-long conversion can usually be done without a problem. The long-to-wide conversion can be difficult. If individuals are seen at different times, direct conversion is impractical. The number of columns in the wide format becomes overly large, and each column contains many missing values. An ad hoc solution is to create homogeneous time groups, which then become the new columns in the wide format. Such regrouping will lead to loss of precision of the time variable. For some studies this need not be a problem, but for others it will.

A more general approach is to impute data in the long format, which requires some form of multilevel imputation. Section 9.3 discusses multiple imputation in the long format. The application defines a common time raster for all persons. Multiple imputations are drawn for each raster point. The resulting imputed datasets can be converted to, and analyzed in, the wide format if desired. This approach is a more principled way to deal with the information loss problem discussed previously. The procedure aligns times to a common raster, hence the name *time raster imputation* (cf. Section 9.3).

9.2 SE Fireworks Disaster Study

On May 13, 2000, a catastrophic fireworks explosion occurred at SE Fireworks in Enschede, the Netherlands. The explosion killed 23 people and injured about 950. Around 500 houses were destroyed, leaving 1,250 people homeless. Ten thousand residents were evacuated.

The disaster marked the starting point of a major operation to recover from the consequences of the explosion. Over the years, the neighborhood has been redesigned and rebuilt. Right after the disaster, the evacuees were relocated to improvised housing. Those in need received urgent medical care. A considerable number of residents showed signed of post-traumatic stress disorder (PTSD). This disorder is associated with flashback memories, avoidance of behaviors, places, or people that might lead to distressing memories, sleeping disorders and emotional numbing. When these symptoms persist and disrupt normal daily functioning, professional treatment is indicated.

Amidst the turmoil in the aftermath, Mediant, the disaster health aftercare center, embedded a randomized controlled trial comparing two treatments for anxiety-related disorders: Eye Movement Desensitization and Reprocessing (EMDR) (Shapiro, 2001) and cognitive behavioral therapy (CBT) (Stallard, 2006). CBT is the standard therapy. The data collection started within one year of the explosion, and lasted until the year 2004 (De Roos et al., 2011). The study included $n = 52$ children 4–18 years, as well as their parents. Children were randomized to EMDR or CBT by a flip of the coin. Each group contained 26 children.

The children received up to four individual sessions over a 4–8 week period, along with up to four parent sessions. Blind assessment took place pre-treatment (T1) and post-treatment (T2) and at 3 months follow-up (T3). The primary outcomes were the UCLA PTSD Reaction Index (PTSD-RI)(Steinberg et al., 2004), the Child Report of Post-traumatic Symptoms (CROPS) and the Parent Report of Post-traumatic Symptoms (PROPS)(Greenwald and Rubin, 1999). Treatment was stopped if children were asymptomatic according to participant and parent verbal report (both conditions), or if there was no remaining distress associated with the trauma memory, as indicated by a self-reported Subjective Units of Disturbance Scale (SUDS) of 0 (EMDR condition only).

The objective of the study was to answer the following questions:

- Is one of these treatments more effective in reducing PTSD symptoms at T2 and T3?
- Does the number of sessions needed to produce the therapeutic effect differ between the treatments?

Table 9.1: SE Fireworks Disaster data. The UCLA PTSD Reaction Index of 52 subjects, children and parents, randomized to EMDR or CBT.

id	trt	pp	Y_1^c	Y_2^c	Y_3^c	Y_1^p	Y_2^p	Y_3^p	id	trt	pp	Y_1^c	Y_2^c	Y_3^c	Y_1^p	Y_2^p	Y_3^p
1	E	Y	–	–	–	36	35	38	32	E	N	28	17	8	40	42	33
2	C	N	45	–	–	–	–	–	33	E	N	–	–	–	38	22	25
3	E	N	–	–	–	13	19	13	34	E	N	–	–	–	17	–	–
4	C	Y	–	–	–	33	27	20	35	E	Y	50	20	–	19	1	5
5	E	Y	26	6	4	27	16	11	37	C	N	30	–	26	59	–	28
6	C	Y	8	1	2	32	15	13	38	C	Y	–	–	–	35	24	27
7	C	Y	41	26	31	–	39	39	39	E	N	–	–	–	–	–	–
8	C	N	–	–	–	24	13	35	40	E	Y	25	5	2	42	13	11
10	C	Y	35	27	14	48	23	–	41	E	Y	36	11	9	30	2	1
12	C	Y	28	15	13	45	33	36	43	E	N	17	–	–	–	–	–
13	E	Y	–	–	–	26	17	14	44	E	N	27	–	–	40	–	–
14	C	Y	33	8	9	37	7	3	45	C	Y	31	12	29	34	28	29
15	E	Y	43	–	7	25	27	1	46	C	Y	–	–	–	44	35	25
16	C	Y	50	8	35	39	21	34	47	C	Y	–	–	–	30	18	14
17	C	Y	31	21	10	32	21	19	48	E	Y	25	18	–	18	17	2
18	E	Y	30	17	16	47	28	34	49	C	N	24	23	16	44	29	34
19	E	Y	29	6	5	20	14	11	50	E	Y	31	13	9	34	18	13
20	E	Y	47	14	22	44	21	25	51	C	Y	–	–	–	52	13	13
21	C	Y	39	12	12	39	5	19	52	C	Y	30	35	28	–	44	50
23	C	Y	14	12	5	29	9	4	53	C	Y	19	33	21	36	21	21
24	E	N	27	–	–	–	–	–	54	C	N	43	–	–	48	–	–
25	E	Y	6	10	5	25	16	16	55	E	Y	64	42	35	44	31	16
28	C	Y	–	2	6	36	17	23	56	C	Y	–	–	–	37	6	9
29	E	Y	23	23	28	23	25	13	57	C	Y	31	12	–	32	26	–
30	E	Y	–	–	–	20	23	12	58	E	Y	–	–	–	49	28	25
31	C	N	15	24	26	33	36	38	59	E	Y	39	7	–	39	7	–

9.2.1 Intention to treat

Table 9.1 contains the outcome data of all subjects. The columns labeled Y_t^c contain the child data, and the columns labeled Y_t^p contain the parent data at time $t = (1, 2, 3)$. Children under the age of 6 years did not fill in the child form, so their scores are missing.

Of the 52 initial participants 14 children (8 EMDR, 6 CBT) did not follow the protocol. The majority (11) of this group did not receive the therapy, but still provided outcome measurements. The three others received therapy, but failed to provide outcome measures. The combined group is labeled as "drop-out," where the other group is called the "completers" or "per-protocol" group. The missing data patterns for both groups can be obtained as:

```
> yvars <- c("yc1", "yc2", "yc3", "yp1", "yp2",
      "yp3")
> md.pattern(fdd[fdd$pp == "Y", yvars])
```

```
   yp2 yp1 yp3 yc1 yc2 yc3
19   1   1   1   1   1   1  0
 1   1   1   1   0   1   1  1
 1   1   1   1   1   0   1  1
 2   1   1   1   1   1   0  1
 2   1   0   1   1   1   1  1
 1   1   1   0   1   1   1  1
 2   1   1   0   1   1   0  2
10   1   1   1   0   0   0  3
     0   2   3  11  11  14 41
> md.pattern(fdd[fdd$pp == "N", yvars])
  yp1 yc1 yp3 yp2 yc3 yc2
3   1   1   1   1   1   1  0
1   1   1   1   0   1   0  2
3   1   0   1   1   0   0  3
2   1   1   0   0   0   0  4
1   1   0   0   0   0   0  5
3   0   1   0   0   0   0  5
1   0   0   0   0   0   0  6
    4   5   7   8  10  11 45
```

The main reason given for dropping out was that the parents were overburdened (8). Other reasons for dropping out were: refusing to talk (1), language problems (1) and a new trauma rising to the forefront (2). One adolescent refused treatment from a therapist not belonging to his own culture (1). One child showed spontaneous recovery before treatment started (1).

Comparison between the 14 drop-outs and the 38 completers regarding presentation at time of initial assessment yielded no significant differences in any of the demographic characteristics or number of traumatic experiences. On the symptom scales, only the mean score of the PROPS was marginally significantly higher for the drop-out group than for the treatment completers ($t = 2.09$, df $= 48$, $p = .04$).

Though these preliminary analyses are comforting, the best way to analyze the data is to the compare participants in the groups to which they were randomized, regardless of whether they received or adhered to the allocated intervention. Formal statistical testing requires random assignment to groups. The ITT principle is widely recommended as the preferred approach to the analysis of clinical trials. DeMets et al. (2007) and White et al. (2011a) provide a balanced discussions of pros and cons the ITT principle.

9.2.2 Imputation model

The major problem of the ITT principle is that some of the data that are needed are missing. Multiple imputation is a natural way to solve this problem, and thus to enable ITT analyses.

A difficulty in setting up the imputation model in the Firework Disaster Data is the large number of outcome variables relative to the number of cases. Even though the analysis of the data in Table 9.1 is already challenging, the real dataset is more complex than this. There are six additional outcome variables (e.g., the Child Behavior Checklist, or CBCL), each measured over time and similarly structured as in Table 9.1. In addition, some of the outcome measures are to be analyzed on both the subscale level and the total score level. For example, the PTSD-RI has three subscales (intrusiveness/numbing/avoidance, fear/anxiety, and disturbances in sleep and concentration and two additional summary measures (Full PTSD and Partial PTSD). All in all, there were 65 variables in data to be analyzed. Of these, 49 variables were incomplete. The total number of cases was 52, so in order to avoid grossly overdetermined models, the predictors of the imputation model should be selected very carefully.

A first strategy for predictor reduction was to preserve all deterministic relations columns in the incomplete data. This was done by passive imputation. For example, let $Y^p_{a,1}$, $Y^p_{b,1}$ and $Y^p_{c,1}$ represent the scores on three subscales of the PTSD parent form administered at T1. Each of these is imputed individually. The total variable Y^p_1 is then imputed by `mice` in a deterministic way as the sum score.

A second strategy to reduce the number of predictors was to leave out other outcomes, measured at other time points. To illustrate this, a subset of the predictor matrix for imputing $Y^p_{a,1}$, $Y^p_{b,1}$ and $Y^p_{c,1}$ is:

```
> vars <- c("ypa1", "ypb1", "ypc1", "ypa2", "ypb2",
      "ypc2", "ypa3", "ypb3", "ypc3")
> pred[vars[1:3], vars]
     ypa1 ypb1 ypc1 ypa2 ypb2 ypc2 ypa3 ypb3 ypc3
ypa1    0    1    1    1    0    0    1    0    0
ypb1    1    0    1    0    1    0    0    1    0
ypc1    1    1    0    0    0    1    0    0    1
```

The conditional distribution $P(Y^p_{a,1}|Y^p_{b,1}, Y^p_{c,1}, Y^p_{a,2}, Y^p_{a,3})$ leaves out the cross-lagged predictors $Y^p_{b,2}$, $Y^p_{c,2}$, $Y^p_{b,3}$ and $Y^p_{c,3}$. The assumption is the cross-lagged predictors are represented by through their non-cross-lagged predictors. Applying this idea consistently throughout the entire 65×65 predictor matrix brings vast reductions of the number of predictors. The largest number of predictors for any incomplete variable was 23, which still leaves degrees of freedom for residual variation.

Specifying a 65×65 predictor matrix by syntax in `R` is tedious and prone to error. I copied the variable names to Microsoft Excel, defined a square matrix of small cells containing zeroes, and used the menu option `Conditional formatting...` to define a cell color if the cell contains a "1." The option `Freeze Panes` was helpful for keeping variable names visible at all times. After filling in the matrix with the appropriate patterns of ones, I exported it to `R` to be used as argument to the `mice()` function. Excel is convenient for setting up large, patterned imputation models.

The imputations were generated as

```
> dry <- mice(fdd, maxit = 0)
> method <- dry$method
> method["yc1"] <- "~I(yca1 + ycb1 + ycc1)"
> method["yc2"] <- "~I(yca2 + ycb2 + ycc2)"
> method["yc3"] <- "~I(yca3 + ycb3 + ycc3)"
> method["yp1"] <- "~I(ypa1 + ypb1 + ypc1)"
> method["yp2"] <- "~I(ypa2 + ypb2 + ypc2)"
> method["yp3"] <- "~I(ypa3 + ypb3 + ypc3)"
> imp <- mice(fdd, pred = pred, meth = method, maxit = 20,
      seed = 54434)
```

9.2.3 Inspecting imputations

For plotting purposes we need to convert the imputed data into long form. In `R` this can be done as follows:

```
> lowi <- complete(imp, "long", inc = TRUE)
> lowi <- data.frame(lowi, cbcl2 = NA, cbin2 = NA,
      cbex2 = NA)
> lolo <- reshape(lowi, idvar = "id", varying = 11:ncol(lowi),
      direction = "long", new.row.names = 1:(nrow(lowi) *
          3), sep = "")
> lolo <- lolo[order(lolo$.imp, lolo$id, lolo$time),
      ]
> row.names(lolo) <- 1:nrow(lolo)
```

This code executes two wide-to-long transformations in succession. The data are imputed in wide format. The call to `complete()` writes the $m+1$ imputed stacked datasets to `lowi`, which stands for "long-wide." The `data.frame()` statement appends three columns to the data with missing CBCL-scores, since the CBCL was not administered at time point 2. The `reshape()` statement interprets everything from column 11 onward as time-varying variables. As long as the variables are labeled consistently, `reshape()` will be smart enough to identify groups of columns that belong together, and stack them in the double-long format `lolo`. Finally, the result is sorted such that the original data with `lolo$.imp==0` are stored as the first block.

Figure 9.1 plots the profiles from 13 subjects with a missing score on Y_1^p, Y_2^p or Y_3^p in Table 9.1. Some profiles are partially imputed. Examples are subjects 7 (missing T1) and 37 (missing T2). Other profiles are missing entirely, and are thus completely imputed. Examples are subjects 2 and 43. Similar plots can be made for other outcomes. In general, the imputed profiles look similar to the completely observed profiles (not shown).

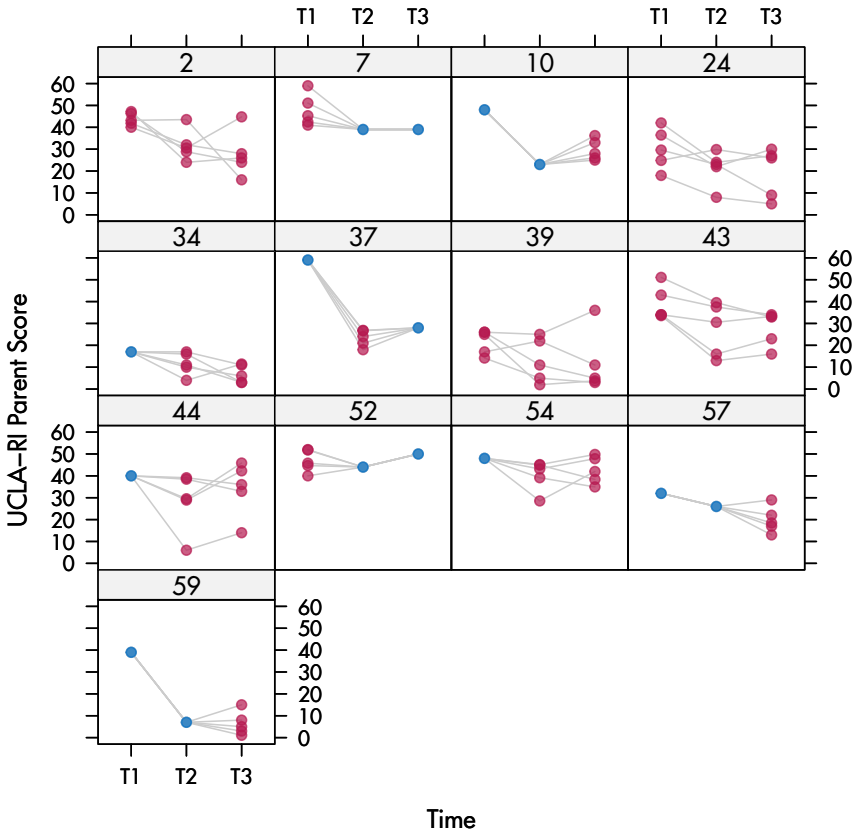


Figure 9.1: Plot of the multiply imputed data of the 13 subjects with one or more missing values on PTSD-RI Parent form.

9.2.4 Complete data analysis

In the absence of missing data, we would have liked to perform a classical repeated measures MANOVA as in Potthoff and Roy (1964). This method construct derived variables that represent time as polynomial contrasts that can be tested. An appealing feature of the method is that the covariances among the repeated measures can take any form.

Let y_{ikt} denote the measurement of individual i ($i = 1, \ldots, n_k$) in group k (CBT or EMDR) at time point t ($t = 1, \ldots, n_t$). In the the SE Firework Disaster Study data, we have $n_k = 26$ and $n_t = 3$. All subjects have been measures at the same time points. The model represents the time trend in

each group by a linear and quadratic trend as

$$y_{ikt} = \beta_{k0} + t\beta_{k1} + t^2\beta_{k2} + e_{ikt} \tag{9.1}$$

where the subject residual e_i has an arbitrary covariance 3×3 matrix Σ that is common to both groups. This model has six β parameters, three for each treatment group. To answer the first research question, we would be interested in testing the null hypotheses $\beta_{11} = \beta_{21}$ and $\beta_{12} = \beta_{22}$, i.e., whether the linear and quadratic trends are different between the treatment groups.

Potthoff and Roy (1964) showed how this model can be transformed into the usual MANOVA model and be fitted by standard software. Suppose that the repeated measures are collected in variables `Y1`, `Y2` and `Y3`. In `SPSS` we can use the `GLM` command to test for the hypothesis of linear and quadratic time trends, and for the hypothesis that these trends are different between CBT and EMDR groups. Though application of the method is straightforward for complete data, it cannot be used directly for the SE Fireworks Disaster data, because of the missing data.

The `mids` object created by `mice()` can be exported as a multiply imputed dataset to `SPSS` by means of the `mids2spss()` function. If the data came originally from `SPSS` it is also possible to merge the imputed data with the original data by means of the `UPDATE` command. `SPSS` will recognize an imported multiply imputed dataset, and execute the analysis m times in parallel. It can also provide the pooled statistics. Note that pooling requires a license to the Missing Values module.

Unfortunately, in `SPSS` 18.0 pooling is not implemented for `GLM`. As a solution, I stored the results by means of the `OMS` command in `SPSS` and shipped the output back to `R` for further analysis. I then applied a yet unpublished procedure for pooling F-tests to the datasets stored by the `OMS` command. In this way, pooling procedures that are not built into `SPSS` can be done with `mice`.

Of course, I could have saved myself the trouble of exporting the imputed data to `SPSS` and performed all analyses in `R`. That would, however, lock out the investigator from her own data. With the new pooling facilities investigators can now do their own data analysis on multiply imputed data. Some re-exporting is therefore worthwhile.

An alternative could have been to create the multiply imputed datasets within `SPSS`. This option was not possible for these data because the `MULTIPLE IMPUTATION` command in `SPSS` does not support predictor selection and passive imputation. With a bit of conversion between software packages, it is possible to have best of both worlds.

9.2.5 Results from the complete data analysis

Figures 9.2 and 9.3 show the development of the mean level of PTSD complaint according to the PTSD-RI. All curves display a strong downward trend between start of treatment (T1) and end of treatment (T2), which is

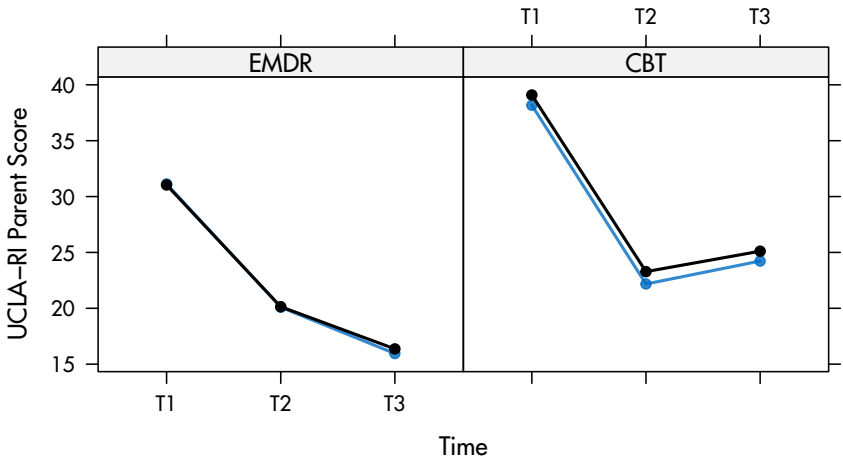


Figure 9.2: Mean levels of PTSD-RI Parent Form for the completely observed profiles (blue) and all profiles (black) in the EMDR and CBT groups.

presumably caused by the EMDR and CBT therapies. The shape between end of treatment (T2) and follow-up (T3) differs somewhat for the group, suggesting that EMDR has better long-term effects, but this difference was not statistically significant. Also note that the complete case analysis and the analysis based on ITT are in close agreement with each other here.

We will not go into details here to answer the second research question as stated on p. 223. It is of interest to note that EMDR needed fewer sessions to achieve its effect. The original publication (De Roos et al., 2011) contains the details.

9.3 Time raster imputation

Longitudinal analysis has become virtually synonymous with mixed effects modeling. Following the influential work of Laird and Ware (1982) and Jennrich and Schluchter (1986), this approach characterizes individual growth trajectories by a small number of random parameters. The differences between individuals are expressed in terms of these parameters.

In some applications, it is natural to consider *change scores*. Change scores are however rather awkward within the context of mixed effects models. This section introduces *time raster imputation*, a new method to generate imputations on a regular time raster from irregularly spaced longitudinal data.

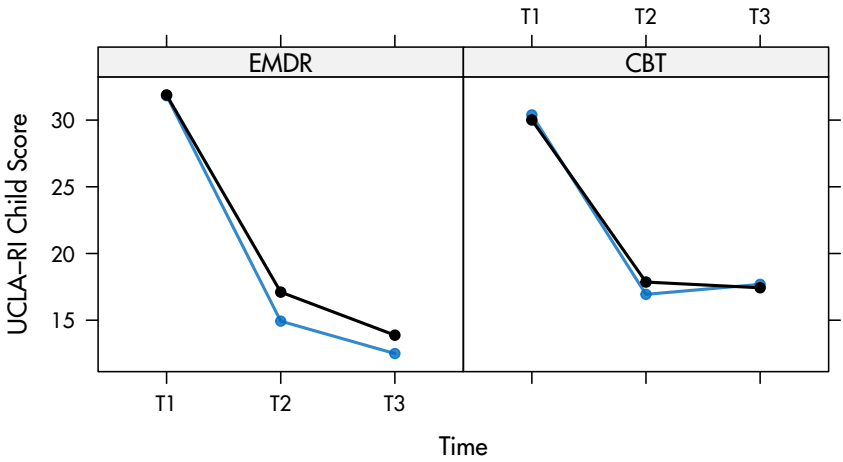


Figure 9.3: Mean levels of PTSD-RI Child Form for the completely observed profiles (blue) and all profiles (black) in the EMDR and CBT groups.

The imputed data can then be used to calculate change scores or age-to-age correlations, or apply quantitative techniques designed for repeated measures.

9.3.1 Change score

Let Y_1 and Y_2 represent repeated measurements of the same object at times T_1 and T_2 where $T_1 < T_2$. The difference $\Delta = Y_2 - Y_1$ is the most direct measure of change over time. Willett (1989, p. 588) characterized the change score as an "intuitive, unbiased, and computationally-simple measure of individual growth."

One would expect that modern books on longitudinal data would take the change score as their starting point. That is not the case. The change score is fully absent from most current books on longitudinal analysis. For example, there is no entry "change score" in the index of Verbeke and Molenberghs (2000), Diggle et al. (2002), Walls and Schafer (2006) or Fitzmaurice et al. (2009). Singer and Willett (2003, p. 10) do discuss the change score, but they quickly dismiss it on the basis that a study with only two time points cannot reveal the shape of a person's growth trajectory.

The change score, once the centerpiece of longitudinal analysis, has disappeared from the methodological literature. I find this is somewhat unfortunate as the parameters in the mixed effects model are more difficult to interpret than the change score. Moreover, classic statistical techniques, like the paired t-test or split-plot ANOVA, are built on the change score. There is a gap be-

tween modern mixed effects models and classical linear techniques for change scores and repeated measures data.

Calculating a mean change score is only sensible if different persons are measured at the same time points. When the data are observed at irregular times, there is no simple way to calculate change scores. Calculating change scores from the person parameters of the mixed effects model is technically trivial, but such scores are difficult to interpret. The person parameters are fitted values that have been smoothed. Deriving a change score as the difference between the fitted curve of the person at T_1 and T_2 results in values that are closer to zero than those derived from data that have been observed.

This section describes a technique that inserts pseudo time points to the observed data of each person. The outcome data at these supplementary time points are multiply imputed. The idea is that the imputed data can be analyzed subsequently by techniques for change scores and repeated measures.

The imputation procedure is akin to the process needed to print a photo in a newspaper. The photo is coded as points on a predefined raster. At the microlevel there could be information loss, but the scenery is essentially unaffected. Hence the name *time raster imputation*. My hope is that this method will help bridge the gap between modern and classic approaches to longitudinal data.

9.3.2 Scientific question: Critical periods

The research was motivated by the question: *At what ages do children become overweight?* Knowing the answer to this question may provide handles for preventive interventions to counter obesity.

Dietz (1994) suggested the existence of three *critical periods* for obesity at adult age: the prenatal period, the period of adiposity rebound (roughly around the age of 5–6 years), and adolescence. Obesity that begins at these periods is expected to increase the risk of persistent obesity and its complications. Overviews of studies on critical periods are given by Cameron and Demerath (2002) and Lloyd et al. (2010).

In the sequel, we use the body mass index (BMI) as a measure of overweight. BMI will be analyzed in standard deviation scores (SDS) using the relevant Dutch references (Fredriks et al., 2000a,b). Our criterion for being overweight in adulthood is defined as BMI SDS ≥ 1.3.

As an example, imagine an 18-year old person with a BMI SDS equal to +1.5 SD. How did this person end up at 1.5 SD? If we have the data, we can plot the measurements against age, and study the individual track. The BMI SDS trajectory may provide key insights into development of overweight and obesity.

Figure 9.4 provides an overview of five theoretical BMI SDS trajectories that the person might have followed. These are:

1. *Long critical period.* A small but persistent centile crossing across the entire age range. In this case, everything (or nothing) is a critical period.

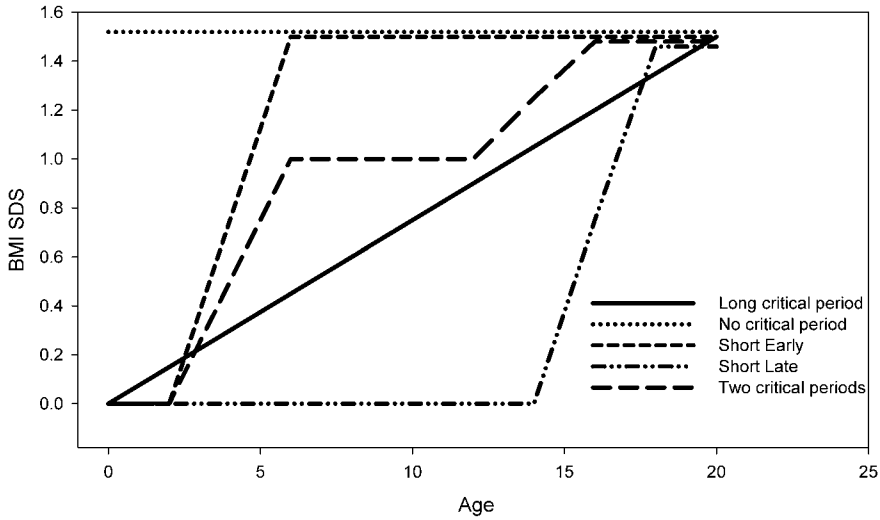


Figure 9.4: Five theoretical BMI SDS trajectories for a person age 18 years with a BMI SDS = 1.5 SD.

2. *No critical period.* The person is born with a BMI SDS of 1.5 SD and this remains unaltered throughout age.

3. *Short early.* There is a large increase between ages 2y and 5y. We would surely interpret the period 2y–5y is as a critical period for this person.

4. *Short late.* This is essentially the same as before, but shifted forward in time.

5. *Two critical periods.* Here the total increase of 1.5 SD is spread over two periods. The first occurs at 2y–5y with an increase of 1.0 SD. The second at 12y–15y with an increase of 0.5 SD.

In practice, mixing between these and other forms will occur.

The objective is to identify any periods during childhood that contribute to an increase in overweight at adult age. A period is "critical" if

1. change differs between those who are and are not later overweight; and

2. change is associated with the outcome after correction for the measure at the end of the period.

Both need to hold. In order to solve the problem of irregular age spacing, De Kroon et al. (2010) use the *broken stick model*, a piecewise linear growth curve fitted, as a means to describe individual growth curves at fixed times.

This section extends this methodology by generating imputations according to the broken stick model. The multiply imputed values are then used to estimate difference scores and regression models that throw light on the question of scientific interest.

9.3.3 Broken stick model ♠

In a sample of n persons $i = 1, \ldots, n$, we assume that there are n_i measurement occasions for person i. Let y_i represent the $n_i \times 1$ vector containing the SDS values obtained for person i. Let t_i represent the $n_i \times 1$ vector with the ages at which the measurements were made.

The broken stick model requires the user to specify an ordered set of k break ages, collected in the vector $\kappa = (\kappa_1, \ldots, \kappa_k)$. The set should cover the entire range of the measured ages, so $\kappa_1 \leq \min(t_i)$ and $\kappa_k \geq \max(t_i)$ for all i. It is convenient to set κ_1 and κ_k to rounded values just below and above the minimum and maximum ages in the data, respectively. De Kroon et al. (2010) specified nine break ages: birth (0d), 8 days (8d), 4 months (4m), 1 year (1y), 2 years (2y), 6 years (6y), 10 years (10y), 18 years (18y) and 29 years (29y).

Without loss of information, the time points t_i of person i are represented by a B-spline of degree 1, with knots specified by κ. More specifically, the vector t_i is recoded as the $n_i \times k$ design matrix $X_i = (x_{1i}, \ldots, x_{ki})$. We refer to Ruppert et al. (2003, p. 59) for further details. For the set of break ages we calculate the B-splines matrix in `R` by the `bs()` function from the `splines` package as follows:

```
> library(splines)
> data <- tbc
> brk <- c(0, 8/365, 1/3, 1, 2, 6, 10, 18, 29)
> k <- length(brk)
> X <- bs(data$age, knots = brk, B = c(brk[1], brk[k] +
      1e-04), degree = 1)
> X <- X[, -(k + 1)]
> dimnames(X)[[2]] <- paste("x", 1:ncol(X), sep = "")
> data <- cbind(data, X)
> round(head(X, 3), 2)
       x1   x2   x3 x4 x5 x6 x7 x8 x9
[1,] 1.00 0.00 0.00  0  0  0  0  0  0
[2,] 0.27 0.73 0.00  0  0  0  0  0  0
[3,] 0.00 0.83 0.17  0  0  0  0  0  0
```

Matrix X has only two nonzero elements in each row. Each row sums to 1. If an observed age coincides with a break age, the corresponding entry is equal to 1, and all remaining elements are zero. In the data example, this occurs in the first record, at birth. A small constant of 0.0001 was added to the last break age. This was done to accomodate for a pseudo time point with an exact age of 29 years, which will be inserted later in Section 9.3.6.

The measurements y_i for person i are modeled by the linear mixed effects model

$$\begin{aligned} y_i &= X_i(\beta + \beta_i) + \epsilon_i \\ &= X_i\gamma_i + \epsilon_i \end{aligned} \tag{9.2}$$

where $\gamma_i = \beta + \beta_i$. The $k \times 1$ column vector β contains k fixed-effect coefficients common to all persons. The vector β_i contains k subject-specific random effect coefficients for person i. The vector ϵ_i contains n_i subject-specific residuals.

We make the usual assumption that $\gamma_i \sim N(\beta, \Omega)$, i.e., the random coefficients of the subjects have a multivariate normal distribution with global mean β and an unstructured covariance Ω. We also assume that the residuals are independently and normally distributed as $\epsilon_i \sim N(0, \sigma^2 I(n_i))$ where σ^2 is a common variance parameter. The covariances between β_i and e_i are assumed to be zero.

Since the rows of the B-spline basis all sum to 1, the intercept is implicit. In fact, one could interpret the model as a special form of the random intercept model, where the intercept is represented by a B-spline rather than by the usual column of ones.

The model prescribes that growth follows a straight line between the break ages. In this application, we are not so much interested in what happens *within* the age interval of each period. Rogosa and Willett (1985) contrasted the analysis of individual differences based on change scores with the analysis of individual differences based on multilevel parameters. They concluded that in general the analysis of change scores is inferior to the parameter approach. The exception is when growth is assumed to follow a straight line within the interval of interest. In that case, the change score approach and the mixed effects model are interchangeable (Rogosa and Willett, 1985, p. 225). The straight line assumption is often reasonable in epidemiological studies if the time interval is short (Hui and Berger, 1983). For extra detail, we could add an extra break age within the interval.

The function `lmer()` from the `lme4` package fits the model. Change scores can be calculated from the fixed and random effects as follows:

```
> library(lme4)
> fit <- lmer(wgt.z~0+x1+x2+x3+x4+x5+x6+x7+x8+x9+
                  (0+x1+x2+x3+x4+x5+x6+x7+x8+x9|id),
                  data=data)
> ### calculate size and increment per person
> tsiz <- t(ranef(fit)$id) + fixef(fit)
> tinc <- diff(tsiz)
> round(head(t(tsiz)),2)
```

The $\hat{\gamma}_i$ estimates are found in the variable `tsiz`. Let $\hat{\delta}_{ik} = \hat{\gamma}_{i,j+1} - \hat{\gamma}_{i,j}$ with $j = 1, \ldots, k-1$ denote the successive differences (or increments) of the

elements in $\hat{\gamma}_i$. These are found in the variable `tinc`. We may interpret $\hat{\delta}_i$ as the expected change scores for person i.

The first criterion for a critical period is that change differs between those who are and are not later overweight. A simple analysis for this criterion is the Student's t-test applied to $\hat{\delta}_{ik}$ for every period k. The correlations between $\hat{\delta}_{ik}$ at successive k were generally higher than 0.5, so we analyzed unconditional change scores (Jones and Spiegelhalter, 2009). The second criterion for a critical period involves fitting two regression models, both of which have final BMI SDS at adulthood, denoted by γ_i^{adult}, as their outcome. The two models are:

$$\gamma_i^{\text{adult}} = \hat{\gamma}_{i,j+1}\zeta_{j+1} + \epsilon_i \tag{9.3}$$

$$\gamma_i^{\text{adult}} = \hat{\gamma}_{i,j+1}\eta_{j+1} + \hat{\gamma}_j\eta_j + \varepsilon_i \tag{9.4}$$

which are fitted for $j = 1, \ldots, k-2$. The parameter of scientific interest is the added value of including η_j.

9.3.4 Terneuzen Birth Cohort

The Terneuzen Birth Cohort consists of all ($n = 2604$) newborns in Terneuzen, the Netherlands, between 1977 and 1986. The most recent measurements were made in the year 2005, so the data spans an age range of 0–29 years. Height and weight were measured throughout this age range. More details on the measurement procedures and the data can be found in De Kroon et al. (2008, 2010).

Suppose the model is fitted to weight SDS. The parameters γ_i can be interpreted as attained weight SDS relative to the reference population. This allows us to represent the observed trajectory of each child in a condensed way by k numbers. The values in $\hat{\gamma}_i$ are the set of most likely weight SDS values at each break age, given all true measurements we have of child i. This implies that if the child has very few measurements, the estimates will be close to the global mean. When taken together, the values $\hat{\gamma}_i$ form the broken stick.

Figure 9.5 displays Weight SDS against age for six selected individuals. Child 1259 has a fairly common pattern. This child starts off near the average, but then steadily declines, apart from a blip around 10 months. Child 2447 is fairly constant, but had a major valley near the age of 4 months, perhaps because of a temporary illness. Child 7019 is also typical. The pattern hovers around the mean. Observe that no data beyond 10 years are available for this child. Child 7460 experienced a substantial change in the height/weight proportions during the first year. Child 7646 was born prematurely with a gestational age of 32 weeks. This individual has an unusually large increase in weight between birth and puberty. Child 8046 is aberrant with an unusually large number of weight measurements around the age of 8 days, but was subsequently not measured for about 1.5 years.

Figure 9.5 also displays the individual broken stick estimates for each outcome as a line. Observe that the model follows the individual data points

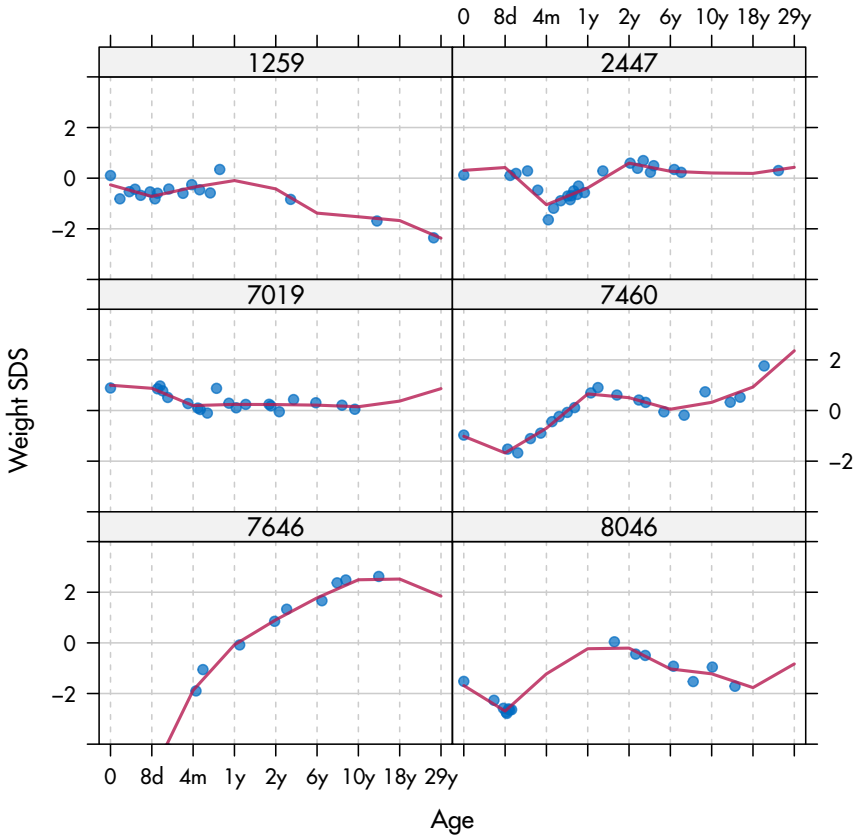


Figure 9.5: Broken stick trajectories for Weight SDS from six selected individuals from the Terneuzen cohort.

very well. De Kroon et al. (2010) analyzed these estimates by the methods described at the end of Section 9.3.2, and found that the periods 2y–6y and 10y–18y were most relevant for developing later overweight.

9.3.5 Shrinkage and the change score ♠

Thus far we have looked at the problem from a prediction perspective. This is a useful first step, but it does not address all aspects. The $\hat{\beta}_i$ estimate in the mixed effects model combines the person-specific ordinary least squares (OLS) estimate of β_i with the grand mean $\hat{\beta}$. The amount of shrinkage toward the grand mean depends on three factors: the number of data points n_i, the residual variance estimate $\hat{\sigma}^2$ around the fitted broken stick, and the variance

estimate $\hat{\omega}_j^2$ for the jth random effect. If $n_i = \sigma^2/\hat{\omega}_j^2$ then $\hat{\beta}_i$ is halfway between $\hat{\beta}$ and the OLS estimate of β_i. If $n_i < \hat{\sigma}^2/\omega_j^2$ then $\hat{\beta}_i$ is closer to the global mean, while $n_i > \hat{\sigma}^2/\hat{\omega}_j^2$ implies that $\hat{\beta}_i$ is closer to the OLS-estimate. We refer to Gelman and Hill (2007, p. 394) for more details.

Shrinkage will stabilize the estimates of persons with few data points. Shrinkage also implies that the same $\hat{\gamma}_i = \hat{\beta} + \hat{\beta}_i$ can correspond to quite different data trajectories. Suppose profile A is an essentially flat and densely measured trajectory just above the mean. Profile B, on the other hand, is a sparse and highly variable trajectory far above the mean. Due to differential shrinkage, profiles A and B can have the same $\hat{\gamma}_i$ estimates. As a consequence, shrinkage will affect the change scores $\hat{\delta}_i$. For both profiles A and B the estimated change scores $\hat{\delta}_i$ are approximately zero at every period. For profile A this is reasonable since the profile itself is flat. In profile B we would expect to see substantial variation in $\hat{\delta}_i$ if the data had been truly measured. Yet, shrinkage has dampened $\hat{\gamma}_i$, and thus made $\hat{\delta}_i$ closer to zero than if calculated from observed data.

It is not quite known whether this effect is a problem in this application. It is likely that dampening of $\hat{\delta}_i$ will bias the result in the conservative direction, and hence primarily affects statistical power. The next section explores an alternative based on multiple imputation. The idea is to insert the break ages into the data, and impute the corresponding outcome data.

9.3.6 Imputation

The measured outcomes are denoted by Y_{obs}, e.g., weight SDS. For the moment, we assume that the Y_{obs} are coded in long format and complete, though neither is an essential requirement. For each person i we append k records, each of which corresponds to a break age. In `R` we use the following statements:

```
> id <- unique(data$id)
> data2 <- appendbreak(data, brk, id = id,
        warp.model = warp.model,  typ = "sup")
> table(data2$typ)
  obs   sup  pred
32845 15705     0
```

The function `appendbreak()` is a custom function of about 20 lines of `R` code specific to the Terneuzen Birth Cohort data. It copies the first available record of the ith person k times, updates administrative and age variables, sets the outcome variables to `NA`, appends the result to the original data and sorts the result with respect to `id` and `age`. The real data are thus mingled with the supplementary records with missing outcomes. The first few records of `data2` look like this:

```
> head(data2)
    id occ nocc first typ   age sex hgt.z  wgt.z  bmi.z ao
4    4   0   19  TRUE obs 0.000   1  0.33  0.195  0.666  0
42   4  NA   19 FALSE sup 0.000   1    NA     NA     NA  0
5    4   1   19 FALSE obs 0.016   1    NA -0.666     NA  0
4.1  4  NA   19 FALSE sup 0.022   1    NA     NA     NA  0
6    4   2   19 FALSE obs 0.076   1  0.71  0.020 -0.381  0
7    4   3   19 FALSE obs 0.104   1  0.18  0.073  0.075  0
      x1   x2   x3 x4 x5 x6 x7 x8 x9 age2
4   1.00 0.00 0.00  0  0  0  0  0  0  0.0
42  1.00 0.00 0.00  0  0  0  0  0  0  0.0
5   0.27 0.73 0.00  0  0  0  0  0  0  2.6
4.1 0.00 1.00 0.00  0  0  0  0  0  0  3.6
6   0.00 0.83 0.17  0  0  0  0  0  0  4.3
7   0.00 0.74 0.26  0  0  0  0  0  0  4.6
```

Multiple imputation must take into account that the data are clustered within persons. The setup for `mice()` requires some care, so we discuss each step in detail.

```
> Y <- c("hgt.z", "wgt.z", "bmi.z")
> imp <- mice(data2, maxit = 0)
> meth <- imp$method
> meth[1:length(meth)] <- ""
> mice.impute.2l.norm.noint <- mice.impute.2l.norm
> meth[Y] <- "2l.norm.noint"
```

These statements specify that only `hgt.z`, `wgt.z` and `bmi.z` need to be imputed. For these three outcomes we request the elementary imputation function `mice.impute.2l.norm()`, which is designed to impute data with two levels. See Section 3.8 for more details.

```
> pred <- imp$pred
> pred[1:nrow(pred), 1:ncol(pred)] <- 0
> pred[Y, "id"] <- (-2)
> pred[Y, "sex"] <- 1
> pred[Y, paste("x", 1:9, sep = "")] <- 2
> pred[Y[1], Y[2]] <- 2
> pred[Y[2], Y[1]] <- 2
> pred[Y[3], Y[1:2]] <- 2
```

The setup of the predictor matrix needs some care. We first empty all entries from the variable `pred`. The statement `pred[Y,"id"] <- (-2)` defines variable `id` as the class variable. The statement `pred[Y,"sex"] <- 1` specifies `sex` as a fixed effect, as usual, while `pred[Y,paste("x",1:9,sep="")] <- 2` sets the B-spline basis as a random effect, as in Equation 9.2. The remaining

three statement specify the Y_2 is a random effects predictor of Y_1 (and vice versa), and both Y_1 and Y_2 are random effects predictors of Y_3. Note that Y_3 (BMI SDS) is not a predictor of Y_1 or Y_2 in order to prevent the type of convergence problems explained in Section 5.5.2. Note also that age is not included in order to evade duplication with its B-spline coding. In summary, there are 12 random effects (9 for age and 3 for the outcomes), one class variable, and one fixed effect.

The actual imputations are produced by

```
> imp.1745.1 <- mice(data2, meth = meth, pred = pred,
      m = 5, maxit = 10, seed = 52711)
> imp.1745.2 <- mice(data2, meth = meth, pred = pred,
      m = 5, maxit = 10, seed = 88348)
> imp.1745 <- ibind(imp.1745.1, imp.1745.2)
> store(imp.1745)
```

When taken together, the calls to `mice()` take about 10 hours. This is much longer than the other applications discussed in this book. The multilevel part of the imputation algorithm runs 300 Gibbs samplers for multilevel analysis on 1745 groups. The two solutions were combined with the `ibind()` function.

Figure 9.6 displays ten multiply imputed trajectories for the six persons displayed in Figure 9.5. The general impression is that the imputed trajectory follows the data quite well. At ages where the are many data points (e.g., in period 0d–1y in person 1259 or in period 8d–1y in person 7460) the curves are quite close, indicating a relatively large certainty. On the other hand, at locations where data are sparse (e.g., the period 10y–29y in person 7019, or the period 8d–2y in person 8046) the curves diverge, indicating a large amount of uncertainty about the imputation. This effect is especially strong at the edges of the age range. Incidentally, we noted that the end effects are less pronounced for larger sample sizes.

It is also interesting to study whether imputation preserves the relation between height, weight and BMI. Figure 9.7 is a scattergram of height SDS and weight SDS split according to age that superposes the imputations on the observed data in the period after the break point. In general the relation in the observed data is preserved in the imputed data. Note that the imputations become more variable for regions with fewer data. This is especially visible at the panel in the upper-right corner at age 29y, where there were no data at all. Similar plots can be made in combination with BMI SDS. In general, the data in these plots all behave as one would expect.

9.3.7 Complete data analysis

Table 9.2 provides a comparison of the mean changes observed under the broken stick model and under time raster imputation. The estimates are very similar, so the mean change estimated under both methods is similar. The p-values in the broken stick method are generally more optimistic relative to

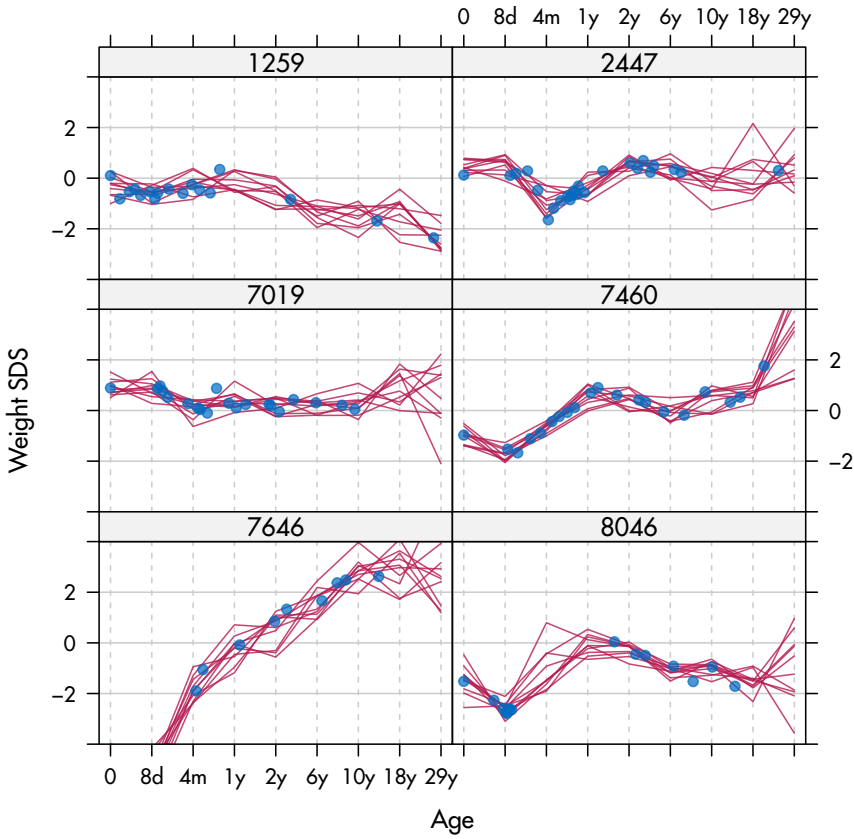


Figure 9.6: Ten multiply imputed trajectories of weight SDS for the same persons as in Figure 9.5 (in red). Also shown are the data points (in blue).

multiple imputation, which is due to the fact that the broken stick model ignores the uncertainty about the estimates.

There is also an effect on the correlations. In general, the age-to-age correlations of the broken stick method are higher than the raster imputations.

Table 9.3 provides the age-to-age correlation matrix of BMI SDS estimated from 1745 cases from the Terneuzen Birth Cohort. Apart from the peculiar values for the age of 8 days, the correlations decrease as the period between time points increases. The values for the broken stick method are higher because these do not incorporate the uncertainty of the estimates.

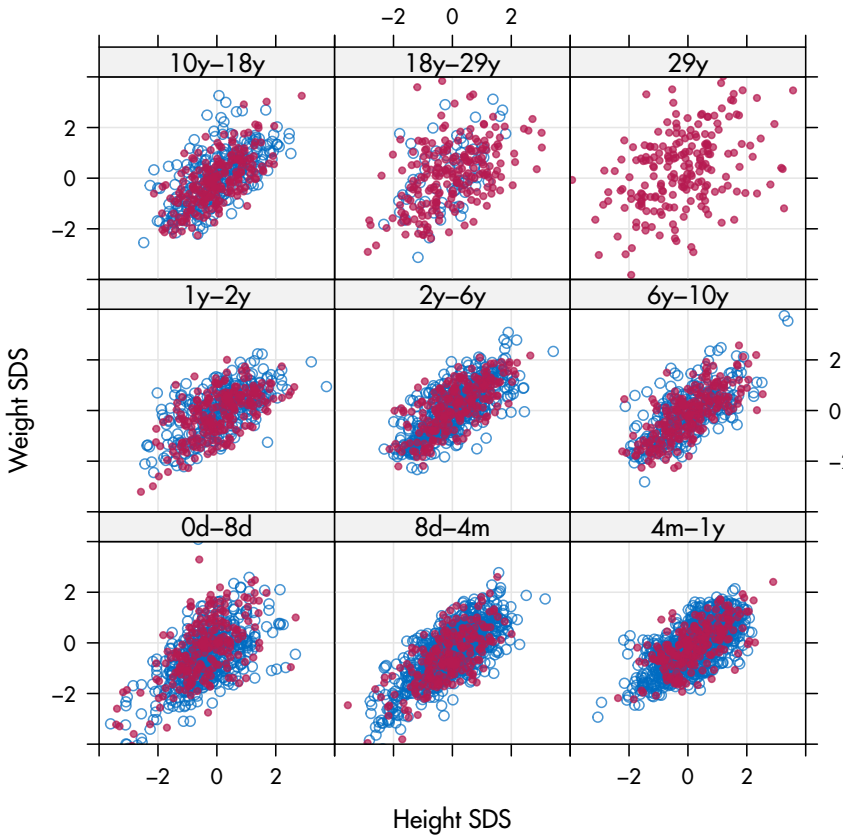


Figure 9.7: The relation between height SDS and weight SDS in the observed (blue) and imputed (red) longitudinal trajectories. The imputed data occur exactly at the break ages. The observed data come from the period immediately after the break age. No data beyond 29 years were observed, so the upper-right panel contains no observed data.

9.4 Conclusion

This chapter described techniques for imputing longitudinal data in both the wide and long formats. Some things are easier in the wide format, e.g., change scores or imputing data, while other procedures are easier in the long format, e.g., graphics and advanced statistical modeling. It is therefore useful to have both formats available.

Table 9.2: Mean change per period, split according to adult overweight (AO) ($n = 124$) and no adult overweight (NAO) ($n = 486$) for the broken stick method and for multiple imputation of the time raster.

Period	Broken stick		*p-value*	Time raster imputation		*p-value*
	NAO	*AO*		*NAO*	*AO*	
0d–8d	−0.88	−0.80	0.214	−0.93	−0.82	0.335
8d–4m	−0.32	−0.34	0.811	−0.07	−0.11	0.745
4m–1y	0.42	0.62	0.006	0.35	0.58	0.074
1y–2y	0.22	0.28	0.242	0.24	0.26	0.884
2y–6y	−0.36	−0.10	<0.001	−0.35	−0.06	0.026
6y–10y	0.05	0.34	<0.001	−0.01	0.31	0.029
10y–18y	0.09	0.52	<0.001	0.17	0.68	0.009

Table 9.3: Age-to-age correlations of BMI SDS the broken stick estimates (lower triangle) and raster imputations (upper triangle) for the Terneuzen Birth Cohort ($n = 1745$).

Age	0d	8d	4m	1y	2y	6y	10y	18y
0d	–	0.64	0.20	0.21	0.18	0.17	0.16	0.11
8d	0.75	–	0.30	0.17	0.20	0.20	0.15	0.13
4m	0.28	0.44	–	0.39	0.30	0.29	0.20	0.16
1y	0.28	0.23	0.65	–	0.55	0.40	0.31	0.23
2y	0.31	0.33	0.46	0.76	–	0.56	0.36	0.23
6y	0.31	0.36	0.46	0.59	0.79	–	0.62	0.42
10y	0.26	0.26	0.35	0.47	0.55	0.89	–	0.53
18y	0.23	0.26	0.29	0.37	0.40	0.72	0.89	–

The methodology for imputing data in the wide format is not really different from that of cross-sectional data. When possible, always try to convert the data into the wide format before imputation. If the data have been observed at irregular time points, as in the Terneuzen Birth Cohort, conversion of the data into the wide format is not possible, however, and imputation can be done in the long format by multilevel imputation.

This chapter introduced time raster imputation, a technique for converting data with an irregular age spacing into the wide format by means of imputation. Time rastering seems to work well in the sense that the generated trajectories follow the individual trajectories. The technique is still experimental and may need further refinement before it can be used routinely.

The current method inserts missing data at the full time grid, and thus imputes data even at time points where there are real observations. One obvious improvement would be to strip such points from the grid so that they are not imputed. For example, in the Terneuzen Birth Cohort this means that we would always take observed birth weight when it is measured.

Another potential improvement is to use the OLS estimates within each

cluster as the center of the posterior predictive distribution rather than their shrunken versions. This would decrease within cluster variability in the imputations, and increase between cluster variability. It is not yet clear how to deal with clusters with only a few time points, but this modification is likely to produce age-to-age correlations that are most faithful to the data.

Finally, the selection of the data could be much stricter. The analysis of the Terneuzen Birth Cohort data used a very liberal inclusion criterion that requires a minimum of only three data points across the entire age range. Sparse trajectories will have large imputation variances, and may thus bias the age-to-age correlations toward zero. As a preliminary rule of thumb, there should be at least one, and preferably two or more, measurements per period.

9.5 Exercises

1. *Potthoff–Roy, wide format imputation.* Potthoff and Roy (1964) published classic data on a study in 16 boys and 11 girls, who at ages 8, 10, 12, and 14 had the distance (mm) from the center of the pituitary gland to the pteryomaxillary fissure measured. Changes in pituitary-pteryomaxillary distances during growth is important in orthodontic therapy. The goals of the study were to describe the distance in boys and girls as simple functions of age, and then to compare the functions for boys and girls. The data have been reanalyzed by many authors including Jennrich and Schluchter (1986), Little and Rubin (1987), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007).

    - Take the version from Little and Rubin (1987) in which nine entries have been made missing. The missing data have been created such that children with a low value at age 8 are more likely to have a missing value at age 10. Use `mice()` to impute the missing entries under the normal model using $m = 100$.
    - For each missing entry, summarize the distribution of the 100 imputations. Determine the interquartile range of each distribution. If the imputations fit the data, how many of the original values you expect to fall within this range? How many actually do?
    - Produce a `lattice` graph of the nine imputed trajectories that clearly shows the range of the imputed values.

2. *Potthoff–Roy comparison.* Use the multiply imputed data from the previous exercise, and apply a linear mixed effects model with an unstructured mean and an unstructured covariance. See Molenberghs and Kenward

(2007, ch. 5) for a discussion of the setup. Discuss advantages and disadvantages of the analysis of the multiply imputed data compared to direct likelihood.

3. *Potthoff–Roy, long format imputation.* Do this exercise with the complete Potthoff–Roy data. Warning: This exercise requires good data handling skills and some patience.

   - Calculate the broken stick estimates for each child using 8, 10, 12 and 14 as the break ages. Make a graph like Figure 9.5. Each data point has exactly one parameter, so the fit could be perfect in principle. Why doesn't that happen? Which two children show the largest discrepancies between the data and the model?
   - Compare the age-to-age correlation matrix of the broken stick estimates to the original data. Why are these correlation matrices different?
   - How would you adapt the analysis such that the age-to-age correlation matrix of the broken stick estimates would reproduce the age-to-age correlation matrix of the original data. Hint: Think of a simpler form of multilevel analysis.
   - Multiply impute the data according to the method used in Section 9.3.6, and produce a display like Figure 9.6 for children 1, 7, 20, 21, 22 and 24.
   - Compare the age-to-age correlation matrix from the imputed data to that of the original data. Are these different? How? Calculate the correlation matrix after deleting the data from the two children who showed the largest discrepancy in the broken stick model. Did this help?
   - How would you adapt the imputation method for the longitudinal data so that its correlation matrix is close to that of the original?

This page intentionally left blank

Part III

Extensions

This page intentionally left blank

Chapter 10

Conclusion

This closing chapter starts with a description of the limitations and pitfalls of multiple imputation. Section 10.2 provides reporting guidelines for applications. Section 10.3 gives an overview of applications that were omitted from the book. Section 10.4 contains some speculations about possible future developments.

10.1 Some dangers, some do's and some don'ts

Any statistical technique has limitations and pitfalls, and multiple imputation is no exception. This books emphasizes the virtues of being flexible, but this comes at a price. The next sections outline some dangers, do's and don'ts.

10.1.1 Some dangers

The major danger of the technique is that it may provide nonsensical or even misleading results if applied without appropriate care or insight. Multiple imputation is not a simple technical fix for the missing data. Scientific and statistical judgment comes into play at various stages: during diagnosis of the missing data problem, in the setup of a good imputation model, during validation of the quality of the generated synthetic data and in combining the repeated analyses. While software producers attempt to set defaults that will work in a large variety of cases, we cannot simply hand over our scientific decisions to the software. We need to open the black box, and adjust the process when appropriate.

The MICE algorithm is univariate optimal, but not necessarily multivariate optimal. There is no clear theoretical rationale for convergence of the multivariate algorithm. The main justification of the MICE algorithm rests on simulation studies. The research on this topic is intensifying. Even though the results obtained thus far are reassuring, at this moment it is not possible to outline in advance the precise conditions that would guarantee convergence for some set of conditionally specified models.

Another danger occurs if the imputation model is uncongenial (Meng, 1994; Schafer, 2003). Uncongeniality can occur if the imputation model is specified as more restrictive than the complete data model, or if it fails to account for important factors in the missing data mechanism. Both types of omissions introduce biased and possibly inefficient estimates. The other side of the coin is that multiple imputation can be more efficient if the imputer uses information that is not accessible to the analyst. The statistical infererences may become more precise than those in maximum likelihood, a property known as *superefficiency* (Rubin, 1996).

There are many data-analytic situations for which we do not yet know the appropriate way to generate imputations. For example, it is not yet clear how design factors of a complex sampling design, e.g., a stratified cluster sample, should be incorporated into the imputation model. Also, relatively little is known about how to impute nested and hierarchical data, or autocorrelated data that form time series. These problems are not inherent limitations of multiple imputation, but of course they may impede the practical application of the imputation techniques for certain types of data.

10.1.2 Some do's

Constructing good imputation models requires analytic skills. The following list of do's summarizes some of the advice given in this book.

- Find out the reasons for the missing data;
- Include the outcome variable(s) in the imputation model;
- Include factors that govern the missingness in the imputation model;
- Impute categorical data by techniques for categorical data;
- Remove response indicators from the imputation model;
- Aim for a scope broad enough for all analyses;
- Set the random seed to enhance reproducible results;
- Break any direct feedback loops that arise in passive imputation;
- Inspect the trace lines for slow convergence;
- Inspect the imputed data;
- Evaluate whether the imputed data could have been real data if they had not been missing;
- Take $m = 5$ for model building, and increase afterward if needed;
- Specify simple MNAR models for sensitivity analysis;

- Impute by proper imputation methods;
- Impute by robust hot deck-like models like predictive mean matching;
- Reduce a large imputation model into smaller components;
- Transform statistics toward approximate normality before pooling;
- Assess critical assumptions about the missing data mechanism;
- Eliminate badly connected, uninteresting variables (low influx, low outflux) from the imputation model;
- Take obvious features like non-negativity, functional relations and skewed distributions in the data into account in the imputation model;
- Use more flexible (e.g., nonlinear or nonparametric) imputation models;
- Perform and pool the repeated analyses per dataset;
- Describe potential departures from MAR;
- Report accurately and concisely.

10.1.3 Some don'ts

Do not:

- Use multiple imputation if simpler methods are valid;
- Take predictions as imputations;
- Impute blindly;
- Put too much faith in the defaults;
- Average the multiply imputed data;
- Create imputations using a model that is more restrictive than needed;
- Uncritically accept imputations that are very different from the observed data.

10.2 Reporting

Section 1.1.2 noted that the attitude toward missing data is changing. Many aspects related to missing data could potentially affect the conclusions drawn for the statistical analysis, but not all aspects are equally important.

This leads to the question: *What should be reported from an analysis with missing data?*

Guidelines to report the results of a missing data analysis have been given by Sterne et al. (2009), Enders (2010), National Research Council (2010) and Mackinnon (2010). These sources vary in scope and comprehensiveness, but they also exhibit a great deal of overlap and consensus. Section 10.2.1 combines some of the material found in the three sources.

Reviewers or editors may be unfamiliar with, or suspicious of, newer approaches to handling missing data. Substantive researchers are therefore often wary about using advanced statistical methods in their reports. Though this concern is understandable,

> ... resorting to flawed procedures in order to avoid criticism from an uninformed reviewer or editor is a poor reason for avoiding sophisticated missing data methodology (Enders, 2010, p. 340)

Until reviewers and referees become more familiar with the newer methods, a better approach is to add well-chosen and concise explanatory notes. On the other hand, editors and reviewers are increasingly expecting applied researchers to do multiple imputation, even when the authors had good reasons for not doing it (e.g., less than 5% incomplete cases) (Ian White, personal communication).

The natural place to report about the missing data in a manuscript is the paragraph on the statistical methodology. As scientific articles are often subject to severe space constraints, part of the report may need to go into supplementary online materials instead of the main text. Since the addition of explanatory notes increases the number of words, there needs to be some balance between the material that goes into the main text and the supplementary material. In applications that requires novel methods, a separate paper may need to be written by the team's statistician. For example, Van Buuren et al. (1999) explained the imputation methodology used in the substantive paper by Boshuizen et al. (1998). In general, the severity of the missing data problem and the method used to deal with the problem needs to be part of the main paper, whereas the precise modeling details could be relegated to the appendix or to a separate methodological paper.

10.2.1 Reporting guidelines

The following list contains questions that need to be answered when using multiple imputation. Evaluate each question carefully, and report the answers.

1. *Amount of missing data:* What is the number of missing values for each variable of interest? What is the number of cases with complete data for the analyses of interest? If people drop out at various time points, break down the number of participants per occasion.

2. *Reasons for missingness:* What is known about the reasons for missing data? Are the missing data intentional? Are the reasons possibly related to the outcome measurements? Are the reasons related to other variables in the study?

3. *Consequences:* Are there important differences between individuals with complete and incomplete data? Do these groups differ in mean or spread on the key variables? What are the consequences if complete case analysis is used?

4. *Method:* What method is used to account for missing data (e.g., complete case analysis, multiple imputation)? Which assumptions were made (e.g., missing at random)? How were multivariate missing data handled?

5. *Software:* What multiple imputation software is used? Which settings differ from the default?

6. *Number of imputed datasets:* How many imputed datasets were created and analyzed?

7. *Imputation model:* Which variables were included in the imputation model? Was any form of automatic variable predictor used? How were non-normally distributed and categorical variables imputed? How were design features (e.g., hierarchical data, complex samples, sampling weights) taken into account?

8. *Derived variables:* How were derived variables (transformations, recodes, indices, interaction terms, and so on) taken into account?

9. *Diagnostics:* How has convergence been monitored? How do the observed and imputed data compare? Are imputations plausible in the sense that they could have been plausibly measured if they had not been missing?

10. *Pooling:* How have the repeated estimates been combined (pooled) into the final estimates? Have any statistics been transformed for pooling?

11. *Complete case analysis:* Do multiple imputation and complete case analysis lead to similar similar conclusions? If not, what might explain the difference?

12. *Sensitivity analysis:* Do the variables included in the imputation model make the missing at random assumption plausible? Are the conclusions affected if imputations are generated under a plausible nonignorable model?

If space is limited, the main text can be restricted to a short summary of points 1, 2, 4, 5, 6 and 11, whereas the remaining points are addressed in a appendix or online supplement. Section 10.2.2 contains an example template.

For clinical trials, reporting in the main text should be extended by point 12, conform to recommendation 15 of National Research Council (2010). Moreover, the study protocol should specify the statistical methods for handling missing data in advance, and their associated assumptions should be stated in a way that can be understood by clinicians (National Research Council, 2010, recommendation 9).

10.2.2 Template

Enders (2010, pp. 340–343) provides four useful templates for reporting the results of a missing data analysis. These templates include explanatory notes for uninformed editors and reviewers. It is straightforward to adapt the template text to other settings. Below I provide a template loosely styled after Enders that I believe captures the essentials needed to report multiple imputation in the statistical paragraph of the main text.

> The percentage of missing values across the nine variables varied between 0 and 34%. In total 1601 out of 3801 records (42%) were incomplete. Many girls had no score because the nurse felt that the measurement was "unnecessary," or because the girl did not give permission. Older girls had many more missing data. We used multiple imputation (Rubin, 1987a) to create and analyze 40 multiply imputed datasets. Methodologists currently regard multiple imputation as a state-of-the-art technique because it improves accuracy and statistical power relative to other missing data techniques. Incomplete variables were imputed under fully conditional specification (Van Buuren et al., 2006). Calculations were done in R 2.13.1 using the default settings of the `mice` 2.12 package (Van Buuren and Groothuis-Oudshoorn, 2011). Model parameters were estimated with multiple regression applied to each imputed dataset separately. These estimates and their standard errors were combined using Rubin's rules. For comparison, we also performed the analysis on the subset of complete cases.

This text is about 150 words. If this is too long, then the sentences that begin with "Methodologists" and "For comparison" can be deleted. In the paragraphs that describe the results we can add the following sentence:

> Table 1 gives the missing data rates of each variable.

In addition, if complete case analysis is included, then we need to summarize it. For example:

> We obtained similar results when the analysis was restricted to the complete cases only. Multiple imputation was generally more efficient as can be seen from the shorter confidence intervals and lower p-values in Table X.

It is also possible that the two analyses lead to diametrically opposed conclusions. Since a well-executed multiple imputation is theoretically superior to complete case analysis, we should give multiple imputation more weight. It would be comforting though to have an explanation of the discrepancy.

The template texts can be adapted as needed. In addition obtain inspiration from good articles in your own field that apply multiple imputation.

10.3 Other applications

Chapters 7–9 illustrated several applications of multiple imputation. This section briefly reviews some other applications. These underscore the general nature and broad applicability of multiple imputation.

10.3.1 Synthetic datasets for data protection

Many governmental agencies make microdata available to the public. One of the major practical issues is that the identity of anonymous respondents can be disclosed through the data they provide. Rubin (1993) suggested publishing fully synthetic microdata instead of the real data, with the obvious advantage of zero disclosure risk. The released synthetic data should reproduce the essential features of confidential microdata.

Raghunathan et al. (2003) and Reiter (2005a) have demonstrated the practical application of the idea. Real and synthetic records can be mixed, resulting in partially synthetic data. Recent work is available as Reiter (2008), Reiter (2009) and Templ (2009).

10.3.2 Imputation of potential outcomes

The effect of a treatment can be expressed as the difference between the observed outcome and the potential outcome, i.e., the outcome that we would have observed if the unit had been allocated to the alternative treatment. This approach is broadly known as the counterfactual approach to causal inference (Morgan and Winship, 2007), and more precisely as the Rubin causal model. By definition, the outcome on the alternative treatment is always missing, so we cannot calculate the treatment effect directly. This restriction is, however, lifted if the missing potential outcomes are imputed. Jin and Rubin (2008) have put this idea into practice to correct for noncompliance in randomized experiments. Multiple imputation of missing potential outcomes can also be useful to correct for imbalances in observational studies. No such studies seem to have appeared yet.

10.3.3 Analysis of coarsened data

Many datasets contain data that are partially missing. Heitjan and Rubin (1991) proposed a general theory for data coarsening processes that includes rounding, heaping, censoring and missing data as special cases. See also Gill et al. (1997) for a slightly more extended model. Heitjan and Rubin (1990) provided an application where age is misreported, where the amount of misreporting increases with age itself. Such problems with the data can be handled by multiple imputation of true age, given reported age and other personal factors. Heitjan (1993) discussed various other biomedical examples and an application to data from the Stanford Heart Transplantation Program. The use of multiple imputation to deal with coarsened data is attractive, but the number of applications to real data (e.g., Heeringa et al. (2002)) has been rather small to date.

10.3.4 File matching of multiple datasets

Statistical file matching, or data fusion, attempts to integrate two or more datasets with different units observed on common variables. Rubin and Schenker (1986a) considered file matching as a missing data problem, and suggested multiple imputation as a solution. Moriarity and Scheuren (2003) developed modifications that were found to improve the procedure. Further relevant work can be found in the books by Rässler (2002), D'Orazio et al. (2006) and Herzog et al. (2007).

The imputation techniques proposed to date were developed from the multivariate normal model. Application of the MICE algorithm under conditional independence is straightforward. Rässler (2002) compared MICE to several alternatives, and found MICE to work well under normality and conditional independence. If the assumption of conditional independence does not hold, we may bring prior information into MICE by appending a third data file that contains records with data that embody the prior information. Sections 5.5.2 and 7.4.5 put this idea into practice in a different context. This techique can perform file matching for mixed continuous-discrete data under any data coded prior.

10.3.5 Planned missing data for efficient designs

Lengthy questionnaires increase the missing data rate and can make a study expensive. An alternative is to cut up a long questionnaire into separate forms, each of which is considerably shorter than the full version. The split questionnaire design (Raghunathan and Grizzle, 1995) poses certain restrictions on the selection of the forms, thus enabling analysis by multiple imputation. Gelman et al. (1998) provide additional techniques for the related problem of analysis of multiple surveys. The loss of efficiency depends on the strengths of the relations between form and can be compensated for by

a larger initial sample size. Graham et al. (2006) is an excellent introduction to methods based on planned missing data. Additional results can be found in Littvay (2009).

10.3.6 Adjusting for verification bias

Partial verification bias in diagnostic accuracy studies may occur if not all patients are assessed by the reference test (golden standard). Bias occurs if the group of patients is selective, e.g., when only those that score on a previous test are measured. Multiple imputation has been suggested as a way to correct for this bias (Harel and Zhou, 2006; De Groot et al., 2008). The classic Begg-Greenes method may be used only if the missing data mechanism is known and simple. For more complex situations De Groot et al. (2011) "strongly recommended" the use of multiple imputation.

10.3.7 Correcting for measurement error

Various authors have suggested multiple imputation to correct for measurement error (Brownstone and Valletta, 1996; Ghosh-Dastidar and Schafer, 2003; Yucel and Zaslavsky, 2005; Cole et al., 2006; Glickman et al., 2008).

10.4 Future developments

Multiple imputation is not a finished product or algorithm. New applications call for innovative ways to implement the key ideas. This section identifies some areas where further research could be useful.

10.4.1 Derived variables

Section 5.4 describes techniques to generate imputations for interactions, sum scores, quadratic terms and other derived variables. Many datasets contain derived variables of some form. The relations between the variables need to be maintained if imputations are to be plausible. There is, however, not yet a lot of experience about how to ensure consistency, and software options are still limited. New types of derived variables and imputation under constraints will call for new techniques and more flexible software.

10.4.2 Convergence of MICE algorithm

A theoretical weakness of the MICE algorithm is that the conditions under which it converges are unknown. In practice, conditionally specified imputa-

tion models are nearly always incompatible. Yet, this fact does not seem to preclude useful and statistically appropriate imputations, as judged by simulation. This suggest that the conditions under which the Gibbs sampler can provide proper imputations can be relaxed. It is unknown, however, how far the conditions may be relaxed.

10.4.3 Algorithms for blocks and batches

In some applications it is useful to generalize the variable-by-variable scheme of the MICE algorithm to blocks. A block can contain just one variable, but also groups of variables. An imputation model is specified for each block, and the algorithm iterates over the blocks. This enables easier specification of blocks of variables that are structurally related, such as dummy variables, semi-continuous variables, bracketed responses, compositions, item subsets, and so on.

Likewise, it may be useful to define batches, groups of records that form logical entities. For example, batches could consist of different populations, time points, classes, and so on. Imputation models can be defined per batch, and iteration takes place over the batches.

The incorporation of blocks and batches will allow for tremendous flexibility in the specification of imputation models. Such techniques require a keen database administration strategy to ensure that the predictors needed at any point are completed. There is currently no imputation software that implements blocks and batches. Javaras and Van Dyk (2003) proposed algorithms for blocks using joint modeling.

10.4.4 Parallel computation

Multiple imputation is a highly parallel technique. If there are m processors available, it is possible to generate the m imputed datasets, estimate the m complete-data statistics, and store the m results by m independent parallel streams. The overhead needed is minimal since each stream requires the same amount of processor time. If more than m processors are available, a better alternative is to subdivide each stream into several substreams. Huge savings can be obtained in this way (Beddo, 2002). It is possible to perform parallel multiple imputation in `R` or `Stata` through macros. Support for multi-core processing is likely to grow.

10.4.5 Nested imputation

In some applications it can be useful to generate different numbers of imputations for different variables. Rubin (2003) described an application that used fewer imputations for variables that were expensive to impute. Alternatively, we may want to impute a dataset that has already been multiply imputed, for example, to impute some additional variables while preserving the original im-

putations. The technique of using different numbers of imputations is known as nested multiple imputation (Shen, 2000). Nested multiple imputation also has potential applications for modeling different types of missing data (Harel, 2009). There is currently no software that supports nested multiple imputation. Extensions to three-step methods have not yet been developed.

10.4.6 Machine learning for imputation

The last two decades have spawned an enormous flood in innovative algorithms for data processing, classification and pattern recognition. See, for example, MacKay (2003), Shawe-Taylor and Cristianni (2004) and Hastie et al. (2009). Problems with missing data increasingly get attention (Parker, 2010). Computational methods like Troyanskaya et al. (2001) typically impute a single value that is "best" in some sense. Using the principles of multiple imputation, it would be extremely valuable to extend data-driven computational techniques to yield a series of plausible values. The evaluation of such generalizations should be done according to the theory of randomization validity.

10.4.7 Incorporating expert knowledge

Honaker and King (2010) observed that imputations in time series are often implausible in the sense that they are too far away from values that experts would expect. One of their suggestions is to incorporate prior expert knowledge about the missing entries, with the purpose of restricting the imputations to a reasonable range. Each missing entry can even have its own prior. The idea is to create imputations as draws from the combined expert prior and posterior predictive distributions. One strategy would be to draw two random values, one from each distribution, and combine these using a mixture model. The statistical properties of the resulting imputations still have to be sorted out.

10.4.8 Distribution-free pooling rules

Rubin's theory is based on a convenient summary of the sampling distribution by the mean and the variance. There seems to be no intrinsic limitation in multiple imputation that would prevent it from working for more elaborate summaries. Suppose that we summarize the distribution of the parameter estimates for each completed dataset by a dense set of quantiles. As before, there will be within- and between-variability as a result of the sampling and missing data mechanisms, respectively. The problem of how to combine these two types of distribution into the appropriate total distribution has not yet been solved. If we would be able to construct the total distribution, this would permit precise distribution-free statistical inference from incomplete data.

10.4.9 Improved diagnostic techniques

The key problem in multiple imputation is how to generate good imputations. The ultimate criterion of a good imputation is that it is confidence proper with respect to the scientific parameters of interest. Diagnostic methods are intermediate tools to evaluate the plausibility of a set of imputations. Section 5.6 discussed several techniques, but these may be laborious for datasets involving many variables. It would be useful to have informative summary measures that can signal whether "something is wrong" with the imputed data. Multiple measures are likely to be needed, each of which can address a specific aspect of the data.

10.4.10 Building block in modular statistics

Multiple imputation requires a well-defined function of the population data, an adequate missing data mechanism, and an idea of the parameters that will be estimated from the imputed data. The technique is an attempt to separate the missing data problem from the complete-data problem, so that both can be addressed independently. This helps in simplifying statistical analyses that are otherwise difficult to optimize or interpret.

The modular nature of multiple imputation helps our understanding. Aided by the vast computational possibilities, statistical models are becoming more and more complex nowadays, up to the point that the models outgrow the capabilities of our minds. The modular approach to statistics starts from a series of smaller models, each dedicated to a particular task. The main intellectual task is to arrange these models in a sensible way, and to link up the steps to provide an overall solution. Compared to the one-big-model-for-everything approach, the modular strategy may sacrifice some optimality. On the other hand, the analytic results are easier to track, as we can inspect what happened after each step, and thus easier to understand. And that is what matters in the end.

10.5 Exercises

1. *Do's:* Take the list of do's in Section 10.1.2. For each item on the list, answer the following questions:
   (a) Why is it on the list?
   (b) Which is the most relevant section in the book?
   (c) Can you order the list elements from most important to least important?
   (d) What were your reasons for picking the top three?

(e) And why did you pick these bottom three?

(f) Could you make suggestions for new items that should be on the list?

2. *Don'ts:* Repeat the previous exercise for the list of don'ts in Section 10.1.3.

3. *Template:* Adapt the template for the `sleep` data in `mice`. Be sure to include your major assumptions and decisions.

4. *Nesting:* Develop an extension of the `mids` object in `mice` that allows for nested multiple imputation. Try to build upon the existing `mids` object.

This page intentionally left blank

Appendix A

Software

This appendix contains an overview of software for multiple imputation. This list is current as of July 2011, but it is likely to become outdated as the developments in this area are fast. Earlier overviews can be found in Horton and Lipsitz (2001), Horton and Kleinman (2007), Harel and Zhou (2007), Yu et al. (2007) and Drechsler (2011). Consult www.multiple-imputation.com for updates.

A.1 R

The `R` language (`R` Development Core Team, 2011) is a public domain general purpose statistical programming language. The following packages implement multiple imputation in `R`, and are freely available from the Comprehensive R Archive Network (CRAN) at `http://www.R-project.org/`.

- `Amelia` 1.2-18 by James Honaker, Gary King and Matthew Blackwell creates multiple imputations based on the multivariate normal model. Specialities include overimputation (remove observed values and impute) and time series imputation.
- `BaBooN` 0.1-6 by Florian Meinfelder generates multiple imputations by chained equations. The package specializes in predictive mean matching for categorical data, and in imputation in data fusion situations where many records have the same missing data pattern.
- `cat` 0.0-6.2 by Joseph L. Schafer implements multiple imputation of categorical data according to the log-linear model as described in Chapters 7 and 8 of Schafer (1997).
- `Hmisc` 3.8-3 by Frank E. Harrell, Jr. contains several functions to diagnose, create and analyze multiple imputations. The major imputation functions are `transcan()` and `aregImpute()`. These functions can automatically transform the data. The function `fit.mult.impute()` combines analysis and pooling and can read `mids` objects created by `mice`.

- `kmi` 0.3-4 by Arthur Allignol performs a Kaplan–Meier multiple imputation, specifically designed to impute missing censoring times.
- `mice` 2.12 by Stef van Buuren and Karin Groothuis-Oudshoorn contributes the chained equations, or MICE algorithm. The package allows for a flexible setup of the imputation model using a predictor matrix and passive imputation. Most of this book was done in `mice`.
- `mi` 0.09-14 by Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima and Maria Grazia Pittau implements a chained equations approach based on Bayesian regression methods. The software allows detailed examination of the fitted imputation model.
- `MImix` 1.0 by Russell Steele, Naisyin Wang and Adrian Raftery implements a special pooling method using a mixture of normal distributions.
- `mitools` 2.0.1 by Thomas Lumley provides tools for analyzing and combining results from multiply imputed data.
- `MissingDataGUI` 0.1-1 by Xiaoyue Cheng, Dianne Cook and Heike Hofmann provides numeric and graphical summaries for the missing values from both discrete and continuous variables. The current version uses `norm` and `Hmisc` to generate multiple imputations.
- `missMDA` 1.2 by Francois Husson and Julie Josse contains the function `MIPCA()` that draws multiple imputations from principal components analysis.
- `miP` 1.1 by Paul Brix can read imputed data created by `Amelia`, `mi` and `mice` to visualize several aspects of the missing data.
- `mirf` 1.0 by Yimin Wu, B. Aletta, S. Nonyane and Andrea S. Foulkes provides a function `mirf()` that creates multiple imputations using random forests.
- `mix` 1.0-8 by Joseph L. Schafer implements the imputation methods based on the general location model as described in Chapter 9 of Schafer (1997).
- `norm` 1.0-9.2 by Joseph L. Schafer implements multiple imputation based on the multivariate normal model as described in Chapters 5 and 6 of Schafer (1997).
- `pan` 0.3 by Joseph L. Schafer implements multiple imputation for multivariate panel or clustered data using the linear mixed model (Schafer and Yucel, 2002).
- `VIM` 2.0.2 by Matthias Templ, Andreas Alfons and Alexander Kowarik introduces tools to visualize missing data before imputation. Imputation functions include `hotdeck()` and `irmi()`, both loosely based on a chained equations approach.

- `Zelig` 3.5 by Kosuke Imai, Gary King and Olivia Lau comes with a general `zelig()` function that supports analysis and pooling of multiply imputed data.

Apart from these, there are many package that contain single imputation methods: `arrayImpute`, `ForImp`, `imputation`, `impute`, `imputeMDR`, `mtsdi`, `missForest`, `robCompositions`, `rrcovNA`, `sbgcop`, `SeqKnn` and `yaImpute`. The functions in these packages typically estimate the missing values in some way, rather than taking random draws.

A.2 S-PLUS

- The `S-PLUS` library `S+MissingData` by Schimert et al. (2001) is the most extensive implementation of the techniques described in Schafer (1997). The library has functions to fit the multivariate Gaussian, log-linear and general location models using EM algorithm and data augmentation (DA) algorithms. The DA algorithms also produce multiple imputations. The library builds upon Schafer's code, but in some cases uses different algorithms. For example, the EM algorithm to fit the Gaussian model uses a Cholesky decomposition of the covariance rather than sweeps as in Schafer's `imp.norm()` function.
- `Hmisc` by Frank E. Harrell Jr. is included as one of the standard libraries. See `R` for the relevant functions.

Versions of `norm`, `cat`, `mix`, `pan` and `mice` for older versions of `S-Plus` (versions 3.3 and 4.0) can still be found on the internet.

A.3 Stata

- The `ice` package by Patrick Royston is a user-contributed `Stata` package that provides an elegant implementation of multiple imputation by chained equations (Royston, 2004, 2009).
- `Stata` 11 introduced the new multiple imputation command `mi`. This is a rich implementation of multiple imputation, including useful options for data manipulation.
- `Stata` 12 (StataCorp LP, 2011) extends the functionality of `mi` with the `mi impute chained` command, which essentially brings the functionality of the `ice` package in the `Stata mi` framework.

A.4 SAS

- Since V8.2, `PROC MI` implements multiple imputation by the multivariate normal model. V9.0 adds predictive mean matching and predictor selection for monotone response patterns.
- Since V8.2, `PROC MIANALYZE` takes the results of the complete data analysis per dataset (e.g., by `PROC LOGISTIC; BY _IMPUTATION_;`) and pools the results. V9.0 adds specification of model effects and custom hypotheses of the parameters.
- `IVEware` (Raghunathan et al., 2002) is an `SAS`-callable software application that implements multiple imputation using chained equations called sequential regressions in `IVEware`. The software allows for flexible imputation models, and includes the ability to specify bounds and data transformations. It has a dedicated command for creating fully synthetic datasets, and routines for regression under complex sampling designs.

A.5 SPSS

- Since SPSS17, `MULTIPLE IMPUTATION`, a part of the Missing Values module, supports multiple imputations by chained equations. Imputation and analysis can be done in a largely automatic fashion and are well integrated with the software for complete data analysis.
- In `AMOS V17` it is possible to generate multiple imputations under the large array of models supported by AMOS.
- `tw.sps` is an `SPSS` macro by Joost van Ginkel that implements *two-way imputation*. This macro can also generate multiple imputations, and is geared toward imputing missing questionnaire items (Van Ginkel et al., 2007).

A.6 Other software

- `SOLAS 4.0` is a stand-alone package dedicated to multiple imputation. `SOLAS` implements five univariate methods to generate multiple imputations. `SOLAS` is the only software that implements propensity score

matching. The software implements a noniterative chained equations method. First, those cells that destroy the monotone pattern are imputed, followed by a second set of imputations of the missing data in the monotone part. Automatic pooling is done for complete data statistics. Several pre- and postimputation diagnostic plots are available.

- `Mplus` Version 6 implements routines to generate, analyze and pool multiply imputed data. Multivariate imputations can be created under a joint model based on the variance-covariance matrix (default) or by a form of conditional specification. `Mplus` embeds multiple imputation using an unrestricted imputation model that is specified behind the scenes (called H1 imputation). It is possible to specify a custom imputation model in conjunction with the Bayesian estimator (called H0 imputation).

- `NORM` is a freeware program by Joseph L. Schafer that imputes missing data under the multivariate normal distribution. The latest version is V2.03. The program contains routines to pool parameter estimates.

- `REALCOM-IMPUTE` is an `MLwiN 2.15` macro that generate imputations for data with two-level structures. The imputation model is an extension of the joint modeling approach to mixed numerical and categorical data with multilevel structure.

- `WinMICE` is a freeware program by Gert Jacobusse that implements multiple imputation under the linear mixed model for two-level data using chained equations. It also contains some features from `mice`.

This page intentionally left blank

References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society C*, 57(3):273–291.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). *Statistical Modelling in R*. Oxford University Press, Oxford.

Ake, C. F. (2005). Rounding after multiple imputation with non-binary categorical covariates. In *Proceedings of the SAS Users Group International (SUGI)*, volume 30, paper 112–30, 1–11.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Allan, F. E. and Wishart, J. (1930). A method of estimating the yield of a missing plot in field experiment work. *Journal of Agricultural Science*, 20(3):399–406.

Allison, P. D. (2002). *Missing Data*. Sage, Thousand Oaks, CA.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. In *Proceedings of the SAS Users Group International (SUGI)*, volume 30, paper 113–30, 1–14.

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide*. SAS Press, Cary, NC, 2nd edition.

Allison, T. and Cicchetti, D. (1976). Sleep in mammals: Ecological and constitutional correlates. *Science*, 194(4266):732–734.

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, 53(1):57–74.

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*. Springer, New York.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (2002). Exact and near compatibility of discrete conditional distributions. *Computational Statistics and Data Analysis*, 40(2):231–252.

Arnold, B. C. and Press, S. J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156.

Austin, P. C. (2008). Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: A simulation study. *Journal of Clinical Epidemiology*, 61(10):1009–1017.

Aylward, D. S., Anderson, R. A., and Nelson, T. D. (2010). Approaches to handling missing data within developmental and behavioral pediatric research. *Journal of Developmental & Behavioral Pediatrics*, 31(1):54–60.

Bang, K. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–972.

Bárcena, M. J. and Tusell, F. (2000). Tree-based algorithms for missing data imputation. In Bethlehem, J. G. and Van der Heijden, P. G. M., editors, *COMPSTAT 2000: Proceedings in Computational Statistics*, pages 193–198. Physica-Verlag, Heidelberg, Germany.

Barnard, J. and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1):17–36.

Barnard, J. and Rubin, D. B. (1999). Small sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955.

Bartlett, M. S. (1978). *An Introduction to Stochastic Processes*. Press Syndicate of the University of Cambridge, Cambridge, UK, 3rd edition.

Beaton, A. E. (1964). The use of special matrix operations in statistical calculus. Research Bulletin RB-64-51, Educational Testing Service, Princeton, NJ.

Bebchuk, J. D. and Betensky, R. A. (2000). Multiple imputation for simple estimation of the hazard function based on interval censored data. *Statistics in Medicine*, 19(3):405–419.

Beddo, V. (2002). Applications of Parallel Programming in Statistics. PhD thesis, University of California, Los Angeles.

Belin, T. R., Hu, M. Y., Young, A. S., and Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine*, 18(22):3123–3135.

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26(6):1368–1382.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36(2):192–236.

Bethlehem, J. G. (2002). Weighting adjustments for ignorable nonresponse. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, chapter 18, pages 275–287. John Wiley & Sons, New York.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15(4):651–675.

Boshuizen, H. C., Izaks, G. J., Van Buuren, S., and Ligthart, G. J. (1998). Blood pressure and mortality in elderly people aged 85 and older: Community based study. *British Medical Journal*, 316(7147):1780–1784.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York.

Brand, J. P. L. (1999). Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets. PhD thesis, Erasmus University, Rotterdam.

Brand, J. P. L., Van Buuren, S., Groothuis-Oudshoorn, C. G. M., and Gelsema, E. S. (2003). A toolkit in SAS for the evaluation of multiple imputation methods. *Statistica Neerlandica*, 57(1):36–45.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth Publishing, New York.

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brownstone, D. and Valletta, R. G. (1996). Modeling earnings measurement error: A multiple imputation approach. *Review of Economics and Statistics*, 78(4):705–717.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Sage, Newbury Park, CA.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076.

Burton, A. and Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *British Journal of Cancer*, 91(1):4–8.

Cameron, N. L. and Demerath, E. W. (2002). Critical periods in human growth and their relationship to diseases of aging. *American Journal of Physical Anthropology*, Suppl 35:159–184.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

Chen, H. Y. (2011). Compatibility of conditionally specified models. *Statistics and Probability Letters*, 80(7–8):670–677.

Chen, H. Y., Xie, H., and Qian, Y. (2011). Multiple imputation for missing values through conditional semiparametric odds ratio models. *Biometrics*, 67(3):799–809.

Chen, L. and Sun, J. (2010). A multiple imputation approach to the analysis of interval-censored failure time data with the additive hazards model. *Computational Statistics and Data Analysis*, 54(4):1109–1116.

Chevalier, A. and Fielding, A. (2011). An introduction to anchoring vignettes. *Journal of the Royal Statistical Society A*, 174(3):569–574.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86(413):68–78.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition.

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple imputation for measurement error correction. *International Journal of Epidemiology*, 35:1074–1081.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–1319.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(3):330–351.

Conversano, C. and Cappelli, C. (2003). Missing data incremental imputation through tree based methods. In Härdle, W. and Rönz, B., editors, *COMPSTAT 2002: Proceedings in Computational Statistics*, pages 455–460. Physica-Verlag, Heidelberg, Germany.

Conversano, C. and Siciliano, R. (2009). Incremental tree-based missing data imputation with lexicographic ordering. *Journal of Classification*, 26(3):361–379.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Creel, D. V. and Krotki, K. (2006). Creating imputation classes using classification tree methodology. In *Proceeding of the Joint Statistical Meeting 2006, ASA Section on Survey Research Methods*, pages 2884–2887. American Statistical Association, Alexandria, VA.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies. Strategies for Bayesian Modeling and Sensitivity Analysis.* Chapman & Hall/CRC, Boca Raton, FL.

Dauphinot, V., Wolff, H., Naudin, F., Guéguen, R., Sermet, C., Gaspoz, J.-M., and Kossovsky, M. (2008). New obesity body mass index threshold for self-reported data. *Journal of Epidemiology and Community Health*, 63(2):128–132.

De Groot, J. A. H., Janssen, K. J. M., Zwinderman, A. H., Bossuyt, P. M. M., Reitsma, J. B., and Moons, K. G. M. (2011). Adjusting for partial verification bias in diagnostic accuracy studies: A comparison of methods. *Annals of Epidemiology*, 21(2):139–148.

De Groot, J. A. H., Janssen, K. J. M., Zwinderman, A. H., Moons, K. G. M., and Reitsma, J. B. (2008). Multiple imputation to correct for partial verification bias: A revision of the literature. *Statistics in Medicine*, 27(28):5880–5889.

De Kroon, M. L. A., Renders, C. M., Kuipers, E. C., Van Wouwe, J. P., Van Buuren, S., De Jonge, G. A., and Hirasing, R. A. (2008). Identifying metabolic syndrome without blood tests in young adults: The Terneuzen birth cohort. *European Journal of Public Health*, 18(6):656–660.

De Kroon, M. L. A., Renders, C. M., Van Wouwe, J. P., Van Buuren, S., and Hirasing, R. A. (2010). The Terneuzen birth cohort: BMI changes between 2 and 6 years correlate strongest with adult overweight. *PloS ONE*, 5(2):e9155.

De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008). *International Handbook of Survey Methodology.* Lawrence Erlbaum Associates, New York.

De Roos, C., Greenwald, R., Den Hollander-Gijsman, M., Noorthoorn, E., Van Buuren, S., and De Jong, A. (2011). A randomised comparison of cognitive behavioral therapy (CBT) and eye movement desensitization and reprocessing (EMDR) in disaster-exposed children. *European Journal of Psychotraumatology*, 2:5694.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, Hoboken, NJ.

DeMets, D. L., Cook, T. D., and Roecker, E. (2007). Selected issues in the analysis. In Cook, T. and DeMets, D., editors, *Introduction to Statistical Methods for Clinical Trials*, chapter 11, pages 339–376. Chapman & Hall/CRC, Boca Raton, FL.

Demirtas, H. (2009). Rounding strategies for multiply imputed binary data. *Biometrical Journal*, 51(4):677–688.

Demirtas, H. (2010). A distance-based rounding strategy for post-imputation ordinal data. *Journal of Applied Statistics*, 37(3):489–500.

Demirtas, H., Freels, S. A., and Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78(1):69–84.

Demirtas, H. and Hedeker, D. (2008a). Imputing continuous data under some non-Gaussian distributions. *Statistica Neerlandica*, 62(2):193–205.

Demirtas, H. and Hedeker, D. (2008b). Multiple imputation under power polynomials. *Communications in Statistics—Simulation and Computation*, 37(8):1682–1695.

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

Dempster, A. P. and Rubin, D. B. (1983). Introduction. In *Incomplete Data in Sample Surveys*, volume 2, pages 3–10. Academic Press, New York.

Dietz, W. H. (1994). Critical periods in childhood for the development of obesity. *American Journal of Clinical Nutrition*, 59(5):955–959.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Clarendon Press, Oxford, 2nd edition.

Dillman, D. A., Smyth, J. D., and Melani Christian, L. (2008). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons, New York, 3rd edition.

Dorans, N. J. (2007). Linking scores from multiple health outcome instruments. *Quality of Life Research*, 16(Suppl 1):85–94.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, UK.

Dorey, F. J., Little, R. J. A., and Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, 12(17):1589–1603.

Drechsler, J. (2011). Multiple imputation in practice: A case study using a complex German establishment survey. *Advances in Statistical Analysis*, 95:1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

El Adlouni, S., Favre, A.-C., and Bobée, B. (2006). Comparison of methodologies to assess the convergence of Markov chain Monte Carlo methods. *Computational Statistics and Data Analysis*, 50(10):2685–2701.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

Farhangfar, A., Kurgan, L. A., and Pedrycz, W. (2007). A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 37(5):692–709.

Fay, R. E. (1992). When are inferences from multiple imputation valid? In *ASA 1992 Proceedings of the Survey Research Methods Section*, pages 227–232. American Statistical Association, Alexandria, VA.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91(434):490–498.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Ford, B. L. (1983). An overview of hot-deck procedures. In Madow, W., Olkin, I., and Rubin, D. B., editors, *Incomplete Data in Sample Surveys*, volume 2, chapter 14, pages 185–207. Academic Press, New York.

Fredriks, A. M., Van Buuren, S., Burgmeijer, R. J. F., Meulmeester, J. F., Beuker, R. J., Brugman, E., Roede, M. J., Verloove-Vanhorick, S. P., and Wit, J. M. (2000a). Continuing positive secular growth change in the Netherlands 1955–1997. *Pediatric Research*, 47(3):316–323.

Fredriks, A. M., Van Buuren, S., Wit, J. M., and Verloove-Vanhorick, S. P. (2000b). Body index measurements in 1996–7 compared with 1980. *Archives of Disease in Childhood*, 82(2):107–112.

García-Laencina, P. J., Sancho-Gómez, J.-S., and Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computation & Applications*, 19(2):263–282.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, London, 2nd edition.

Gelman, A. and Hill, J. L. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, UK.

Gelman, A., Jakulin, A., Grazia Pittau, M., and Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443):846–857.

Gelman, A. and Meng, X.-L., editors (2004). *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley & Sons, Chichester, UK.

Gelman, A. and Speed, T. P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society B*, 55(1):185–188.

Geskus, R. B. (2001). Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Statistics in Medicine*, 20(5):795–812.

Ghosh-Dastidar, B. and Schafer, J. L. (2003). Multiple edit/multiple imputation for multivariate continuous data. *Journal of the American Statistical Association*, 98(464):807–817.

Gilks, W. R. (1996). Full conditional distributions. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, chapter 5, pages 75–88. Chapman & Hall, London.

Gill, R. D., Van der Laan, M. L., and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures and counter-examples. In Lin, D.-Y. and Fleming, T. R., editors, *Proceedings of the First Seattle Conference on Biostatistics*, pages 255–294. Springer-Verlag, Berlin.

Glickman, M. E., He, Y., Yucel, R. M., and Zaslavsky, A. M. (2008). Misreporting, missing data, and multiple imputation: Improving accuracy of cancer registry databases. *Chance*, 21(3):55–58.

Glynn, R. J. and Laird, N. M. (1986). Regression estimates and missing data: Complete-case analysis. Technical report, Harvard School of Public Health, Department of Biostatistics.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In Wainer, H., editor, *Drawing Inferences from Self-Selected Samples*, pages 115–142. Springer-Verlag, Berlin.

Glynn, R. J. and Rosner, B. (2004). Multiple imputation to estimate the association between eyes in disease progression with interval-censored data. *Statistics in Medicine*, 23(21):3307–3318.

Goldstein, H., Carpenter, J. R., Kenward, M. G., and Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–179.

Gonzalez, J. M. and Eltinge, J. L. (2007). Multiple matrix sampling: A review. In *ASA 2007 Proceedings of the Section on Survey Research Methods*, pages 3069–3075. American Statistical Association, Alexandria, VA.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65(329):226–256.

Gorber, S. C., Tremblay, M., Moher, D., and Gorber, B. (2007). A comparison of direct vs. self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Reviews*, 8(4):307–326.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Preventive Science*, 8(3):206–213.

Graham, J. W., Taylor, B. J., Olchowski, A. E., and Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, 11(4):323–343.

Greenland, S. and Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142(12):1255–1264.

Greenwald, R. and Rubin, A. (1999). Brief assessment of children's post-traumatic symptoms: Development and preliminary validation of parent and child scales. *Research on Social Work Practice*, 9(1):61–75.

Groothuis-Oudshoorn, C. G. M., Van Buuren, S., and Van Rijckevorsel, J. L. A. (1999). Flexible multiple imputation by chained equations of the AVO-95 survey. Technical Report PG/VGZ/00.045, TNO Prevention and Health, Leiden.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons, New York, 2nd edition.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman & Hall, London.

Harel, O. (2009). The estimation of r^2 and adjusted r^2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics*, 36(10):1109–1118.

Harel, O. and Zhou, X. H. (2006). Multiple imputation for correcting verification bias. *Statistics in Medicine*, 25(22):3769–3786.

Harel, O. and Zhou, X. H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, 26(16):3057–3077.

Harkness, J. A., Van de Vijver, F. J. R., and Mohler, P. P., editors (2002). *Cross-Cultural Survey Methods*. John Wiley & Sons, New York.

Harrell, F. E. (2001). *Regression Modeling Strategies*. Springer-Verlag, New York.

Hastie, T., Tibshirani, R. J., and Friedman, J. (2009). *The Elements of Statistical Learning*. Spinger, New York, 2nd edition.

He, Y. (2006). Missing Data Imputation for Tree-Based Models. PhD thesis, University of California, Los Angeles, CA.

He, Y. and Raghunathan, T. E. (2006). Tukey's *gh* distribution for multiple imputation. *The American Statistician*, 60(3):251–256.

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C. (2001). Dependency networks for inference, collaborative filtering, and data visualisation. *Journal of Machine Learning Research*, 1(1):49–75.

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(4):475–492.

Heeringa, S. G., Little, R. J. A., and Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, chapter 24, pages 357–371. John Wiley & Sons, New York.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419.

Heitjan, D. F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49(4):1099–1109.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society C*, 40(1):13–29.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85(410):304–314.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19(4):2244–2253.

Herzog, T. N. and Rubin, D. B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In Madow, W., Olkin, I., and Rubin, D. B., editors, *Incomplete Data in Sample Surveys*, volume 2, chapter 15, pages 209–245. Academic Press, New York.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linking Techniques*. Springer, New York.

Heymans, M. W., Anema, J. R., Van Buuren, S., Knol, D. L., Van Mechelen, W., and De Vet, H. C. W. (2009). Return to work in a cohort of low back pain patients: Development and validation of a clinical prediction rule. *Journal of Occupational Rehabilitation*, 19(2):155–165.

Heymans, M. W., Van Buuren, S., Knol, D. L., Anema, J. R., Van Mechelen, W., and De Vet, H. C. W. (2010). The prognosis of chronic low back pain is determined by changes in pain and disability in the initial period. *Spine Journal*, 10(10):847–856.

Heymans, M. W., Van Buuren, S., Knol, D. L., Van Mechelen, W., and De Vet, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*, 7:33.

Hille, E. T. M., Elbertse, L., Bennebroek Gravenhorst, J., Brand, R., and Verloove-Vanhorick, S. P. (2005). Nonresponse bias in a follow-up study of 19-year-old adolescents born as preterm infants. *Pediatrics*, 116(5):662–666.

Hille, E. T. M., Weisglas-Kuperus, N., Van Goudoever, J. B., Jacobusse, G. W., Ens-Dokkum, M. H., De Groot, L., Wit, J. M., Geven, W. B., Kok, J. H., De Kleine, M. J. K., Kollée, L. A. A., Mulder, A. L. M., Van Straaten, H. L. M., De Vries, L. S., Van Weissenbruch, M. M., and Verloove-Vanhorick, S. P. (2007). Functional outcomes and participation in young adulthood for very preterm and very low birth weight infants: The Dutch project on preterm and small for gestational age infants at 19 years of age. *Pediatrics*, 120(3):587–595.

Holland, P. W. (2007). A framework and history for score linking. In Dorans, N. J., Pommerich, M., and Holland, P. W., editors, *Linking and aligning scores and scales*, chapter 2, pages 5–30. Springer, New York.

Holland, P. W. and Wainer, H., editors (1993). *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the Arctic. *Biometrics*, 57(1):22–33.

Hopman-Rock, M., Dusseldorp, E., Chorus, A. M. J., Jacobusse, G. W., Rütten, A., and Van Buuren, S. (2012). Response conversion for improving comparability of international physical activity data. *Journal of Physical Activity & Health*, 9(1):29–38.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90.

Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3):244–254.

Horton, N. J., Lipsitz, S. R., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57(4):229–232.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, New York, 2nd edition.

Hosmer, D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Hron, K., Templ, M., and Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. *Computational Statistics and Data Analysis*, 54(12):3095–3107.

Hsu, C. H. (2007). Multiple imputation for interval censored data with auxiliary variables. *Statistics in Medicine*, 26(4):769–781.

Hsu, C. H., Taylor, J. M. G., Murray, S., and Commenges, D. (2006). Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*, 25(20):3503–3517.

Hui, S. L. and Berger, J. O. (1983). Empirical Bayes estimation of rates in longitudinal studies. *Journal of the American Statistical Association*, 78(384):753–760.

Ip, E. H. and Wang, Y. J. (2009). Canonical representation of conditionally specified multivariate discrete distributions. *Journal of Multivariate Analysis*, 100(6):1282–1290.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3):841–860.

Izaks, G. J., Van Houwelingen, H. C., Schreuder, G. M., and Ligthart, G. J. (1997). The association between human leucocyte antigens (HLA) and mortality in community residents aged 85 and older. *Journal of the American Geriatrics Society*, 45(1):56–60.

James, I. R. and Tanner, M. A. (1995). A note on the analysis of censored regression data by multiple imputation. *Biometrics*, 51(1):358–362.

Javaras, K. N. and Van Dyk, D. A. (2003). Multiple imputation for incomplete data with semicontinuous variables. *Journal of the American Statistical Association*, 98(463):703–715.

Jeliĉić, H., Phelps, E., and Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, 45(4):1195–1199.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *Journal of the American Statistical Association*, 103(481):101–111.

Jones, H. E. and Spiegelhalter, D. J. (2009). Accounting for regression-to-the-mean in tests for recent changes in institutional performance: Analysis and power. *Statistics in Medicine*, 30(12):1645–1667.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kasim, R. M. and Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, 23(2):93–116.

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *ASA 1991 Proceedings of the Section on Survey Research Methods*, pages 1–10.

Kenward, M. G. and Molenberghs, G. (2009). Last observation carried forward: A crystal ball? *Journal of Biopharmaceutical Statistics*, 19(5):872–888.

Khare, M., Little, R. J. A., Rubin, D. B., and Schafer, J. L. (1993). Multiple imputation of NHANES III. In *ASA 1993 Proceedings of the Survey Research Methods Section*, volume 1, pages 297–302. American Statistical Association, Alexandria, VA.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society B*, 68(3):509–521.

Kim, J. K. and Fuller, W. A. (2004). Fractional hot deck imputation. *Biometrika*, 91(3):559–578.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69.

King, G., Murray, C. J. L., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1):191–207.

Klebanoff, M. A. and Cole, S. R. (2008). Use of multiple imputation in the epidemiologic literature. *American Journal of Epidemiology*, 168(4):355–357.

Kleinbaum, D. G. and Klein, M. B. (2005). *Survival Analysis: A Self-Learning Text*. Springer-Verlag, New York, 2nd edition.

Knol, M. J., Janssen, K. J. M., Donders, A. R. T., Egberts, A. C. G., Heerdink, E. R., Grobbee, D. E., Moons, K. G. M., and Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology*, 63:728–736.

Kolen, M. J. and Brennan, R. L. (1995). *Test Equating: Methods and Practices*. Springer, New York.

Koller-Meinfelder, F. (2009). Analysis of Incomplete Survey Data—Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching. PhD thesis, Otto-Friedrich-Universität, Bamberg.

Krul, A., Daanen, H. A. M., and Choi, H. (2010). Self-reported and measured weight, height and body mass index (BMI) in Italy, the Netherlands and North America. *European Journal of Public Health*, 21(4):414–419.

Kuo, K.-L. and Wang, Y. J. (2011). A simple algorithm for checking compatibility among discrete conditional distributions. *Computational Statistics and Data Analysis*, 55(8):2457–2462.

Lagaay, A. M., Van der Meij, J. C., and Hijmans, W. (1992). Validation of medical history taking as part of a population based survey in subjects aged 85 and over. *British Medical Journal*, 304(6834):1091–1092.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lam, K. F., Tang, O. Y., and Fong, D. Y. T. (2005). Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine*, 24(12):1865–1879.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896.

Lee, H., Rancourt, E., and Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10(3):231–243.

Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5):624–632.

Lesaffre, E. and Albert, A. (1989). Partial separation in logistic discrimination. *Journal of the Royal Statistical Society B*, 51(1):109–116.

Li, K.-H. (1988). Imputation using Markov chains. *Journal of Statistical Computation and Simulation*, 30(1):57–79.

Li, K.-H., Meng, X.-L., Raghunathan, T. E., and Rubin, D. B. (1991a). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, 1(1):65–92.

Li, K.-H., Raghunathan, T. E., and Rubin, D. B. (1991b). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86(416):1065–1073.

Little, R. J. A. (1988). Missing-data adjustments in large surveys (with discussion). *Journal of Business Economics and Statistics*, 6(3):287–301.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. A. (2009). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 18, pages 409–431. CRC Press, Boca Raton, FL.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2nd edition.

Littvay, L. (2009). Questionnaire design considerations with planned missing data. *Review of Psychology*, 16(2):103–113.

Liu, C. (1993). Barlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *Journal of Multivariate Analysis*, 46(2):198–206.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis*, 53(1):139–158.

Liu, C. and Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika*, 85(3):673–688.

Liu, L. X., Murray, S., and Tsodikov, A. (2011). Multiple imputation based on restricted mean model for censored data. *Statistics in Medicine*, 30(12):1339–1350.

Lloyd, L. J., Langley-Evans, S. C., and McMullen, S. (2010). Childhood obesity and adult cardiovascular disease risk: A systematic review. *International Journal of Obesity*, 34(1):18–28.

Lyles, R. H., Fan, D., and Chuachoowong, R. (2001). Correlation coefficient estimation involving a left censored laboratory assay variable. *Statistics in Medicine*, 20(19):2921–2933.

Lynn, H. S. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*, 20(1):33–45.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

Mackinnon, A. (2010). The use and reporting of multiple imputation in medical research: A review. *Journal of Internal Medicine*, 268(6):586–593.

Madow, W. G., Olkin, I., and Rubin, D. B., editors (1983). *Incomplete Data in Sample Surveys*, volume 2. Academic Press, New York.

Marker, D. A., Judkins, D. R., and Winglee, M. (2002). Large-scale imputation for complex surveys. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, chapter 22, pages 329–341. John Wiley & Sons, New York.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling*, 5(1):22–36.

Marshall, A., Altman, D. G., and Holder, R. L. (2010a). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study research article open access. *BMC Medical Research Methodology*, 10:112.

Marshall, A., Altman, D. G., Royston, P., and Holder, R. L. (2010b). Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Research Methodology*, 10:7.

Marshall, A., Billingham, L. J., and Bryan, S. (2009). Can we afford to ignore missing data in cost-effectiveness analyses? *European Journal of Health Economics*, 10(1):1–3.

Matsumoto, D. and Van de Vijver, F. J. R., editors (2010). *Cross-Cultural Research Methods in Psychology*. Cambridge University Press, Cambridge.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, New York, 2nd edition.

McKnight, P. E., McKnight, K. M., Sidani, S., and Figueredo, A. J. (2007). *Missing Data. A Gentle Introduction*. Guilford Press, New York.

Meng, X.-L. (1994). Multiple imputation with uncongenial sources of input (with discusson). *Statistical Science*, 9(4):538–573.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1):103–111.

Miettinen, O. S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. John Wiley & Sons, New York.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. John Wiley & Sons, Chichester, UK.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

Moons, K. G. M., Donders, A. R. T., Stijnen, T., and Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59(10):1092–1101.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference. Methods and Principles for Social Research*. Cambridge University Press, Cambridge.

Moriarity, C. and Scheuren, F. J. (2003). Note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business Economics and Statistics*, 21(1):65–73.

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, Washington, D.C.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*, 71(3):593–627.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with discrete and continuous variables. *Annals of Mathematical Statistics*, 32(2):448–465.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 697–715.

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56(1):199–203.

Pan, W. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics*, 57(4):1245–1250.

Parker, R. (2010). *Missing Data Problems in Machine Learning*. VDM Verlag Dr. Müller, Saarbrücken, Germany.

Peng, Y., Little, R. J. A., and Raghunathan, T. E. (2004). An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics*, 60(3):598–607.

Peugh, J. L. and Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Spinger, New York.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model usefully especially for growth curve problems. *Biometrika*, 51(3):313–326.

Raghunathan, T. E. and Bondarenko, I. (2007). Diagnostics for multiple imputations. Technical report, Department of Biostatistics, University of Michigan. http://ssrn.com/abstract=1031750.

Raghunathan, T. E. and Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. W. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1):85–95.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software User Guide*. Survey Methodology Program Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91(434):499–505.

Rässler, S. (2002). *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer, New York.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiter, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society A*, 168(1):185–205.

Reiter, J. P. (2005b). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3):7–30.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502–508.

Reiter, J. P. (2008). Selecting the number of imputed datasets when using multiple imputation for missing data and disclosure limitation. *Statistics and Probability Letters*, 78(1):15–20.

Reiter, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review*, 77(2):179–195.

Renn, S. D. (2005). *Expository Dictionary of Bible Words: Word Studies for Key English Bilble Words Based on the Hebrew and Greek Texts*. Hendrickson Publishers, Peabody, MA.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society C*, 54(3):507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2006). Using the Box–Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, 6(3):209–229.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, chapter 3, pages 45–57. Chapman & Hall, London.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1):113–124.

Rogosa, D. R. and Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50(2):203–228.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal*, 4(3):227–241.

Royston, P. (2007). Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal*, 7(4):445–464.

Royston, P. (2009). Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables. *Stata Journal*, 9(3):466–477.

Royston, P., Parmar, M. K. B., and Sylvester, R. (2004). Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Statistics in Medicine*, 23(6):907–926.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–590.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business Economics and Statistics*, 4(1):87–94.

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rubin, D. B. (1987b). A noniterative sample/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fractions of missing information are modest. *Journal of the American Statistical Association*, 82(398):543–546.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468.

Rubin, D. B. (1994). Comments on "Missing data, imputation, and the bootstrap" by Bradley Efron. *Journal of the American Statistical Association*, 89(426):485–488.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57(1):3–18.

Rubin, D. B. (2004). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, 58(4):298–302.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *ASA 1990 Proceedings of the Statistical Computing Section*, pages 83–88, American Statistical Association, Alexandria, VA.

Rubin, D. B. and Schenker, N. (1986a). Efficiently simulating the coverage properties of interval estimates. *Journal of the Royal Statistical Society C*, 35(2):159–167.

Rubin, D. B. and Schenker, N. (1986b). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394):366–374.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657.

Salomon, J. A., Tandon, A., and Murray, C. J. L. (2004). Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes. *British Medical Journal*, 328(7434):258.

Särndal, C. E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, New York.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11(16):2093–2109.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35.

Schafer, J. L., Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., and Rubin, D. B. (1996). The NHANES III multiple imputation project. In *ASA 1996 Proceedings of the Survey Research Methods Section*, pages 28–37, Alexandria, VA.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571.

Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95(449):144–154.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11(2):437–457.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94(448):1096–1120.

Schenker, N., Raghunathan, T. E., and Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, 29(5):533–545.

Schenker, N. and Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, 22(4):425–446.

Scheuren, F. J. (2004). Introduction to history corner. *The American Statistician*, 58(4):290–291.

Scheuren, F. J. (2005). Multiple imputation: How it began and continues. *The American Statistician*, 59(4):315–319.

Schimert, J., Schafer, J. L., Hesterberg, T., Fraley, C., and Clarkson, D. B. (2001). *Analyzing Data with Missing Values in S-PLUS*. Insightful Corporation, Seattle, WA.

Schönbeck, Y., Talma, H., Van Dommelen, P., Bakker, B., Buitendijk, S. E., Hirasing, R. A., and Van Buuren, S. (2011). Increase in prevalence of overweight in Dutch children and adolescents: A comparison of nationwide growth studies in 1980, 1997 and 2009, PLoS ONE, 6(11), e27608.

Scott Long, J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage, Thousand Oaks, CA.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing, Florence, KY, 2nd edition.

Shapiro, F. (2001). *EMDR: Eye Movement Desensitization of Reprocessing: Basic Principles, Protocols and Procedures*. Guilford Press, New York, 2nd edition.

Shawe-Taylor, J. and Cristianni, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Shen, Z. (2000). Nested Multiple Imputation. PhD thesis, Department of Statistics, Harvard University, Cambridge, MA.

Siciliano, R., Aria, M., and D'Ambrosio, A. (2006). Boosted incremental tree-based imputation of missing data. In Zani, S., Cerioli, A., Riani, M., and Vichi, M., editors, *Data Analysis, Classification and the Forward Search*, pages 271–278. Springer, Berlin.

Siddique, J. and Belin, T. R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, 27(1):83–102.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Oxford.

Song, J. and Belin, T. R. (2004). Imputation for incomplete high-dimensional multivariate normal data using a common factor model. *Statistics in Medicine*, 23(18):2827–2843.

Stallard, P. (2006). Psychological interventions for post-traumatic reactions in children and young people: A review of randomised controlled trials. *Clinical Psycholological Review*, 26(7):895–911.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46.

StataCorp LP (2011). *Multiple-Imputation Reference Manual*. StataCorp, College Station, TX.

Steinberg, A. M., Brymer, M. J., Decker, K. B., and Pynoos, R. S. (2004). The University of California at Los Angeles post-traumatic stress disorder reaction index. *Current Psychiatry Reports*, 6(2):96–100.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, 338:b2393.

Steyerberg, E. W. (2009). *Clinical Prediction Models*. Springer, New York.

Subramanian, S. (2009). The multiple imputations based Kaplan–Meier estimator. *Statistics and Probability Letters*, 79(18):1906–1914.

Subramanian, S. (2011). Multiple imputations and the missing censoring indicator model. *Journal of Multivariate Analysis*, 102(1):105–117.

Tan, M. T., Tian, G.-L., and Ng, K. W. (2010). *Bayesian Missing Data Problem. EM, Data Augmentation and Noniterative Computation*. Chapman & Hall/CRC, Boca Raton, FL.

Tang, L., Unüntzer, J., Song, J., and Belin, T. R. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, 24(14):2111–2128.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82(398):528–550.

Taylor, J. M. G., Cooper, K. L., Wei, J. T., Sarma, A. V., Raghunathan, T. E., and Heeringa, S. G. (2002). Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. *American Journal of Epidemiology*, 156(8):774–782.

Tempelman, D. C. G. (2007). Imputation of Restricted Data. PhD thesis, University of Groningen, Groningen.

Templ, M. (2009). New Developments in Statistical Disclosure Control and Imputation. Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken, Germany.

Tian, G.-L., Tan, M. T., Ng, K. W., and Tang, M.-L. (2009). A unified method for checking compatibility and uniqueness for finite discrete conditional distributions. *Communications in Statistics—Theory and Methods*, 38(1):115–129.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, chapter 4, pages 59–74. Chapman & Hall, London.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R. J., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525.

US Bureau of the Census (1957). *United States census of manufactures, 1954, Vol II, Industry Statistics, Part 1, General Summary and Major Groups 20 to 28*. US Bureau of the Census, Washington DC.

Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. Springer-Verlag, Berlin.

Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, 134(8):895–907.

Van Bemmel, T., Gussekloo, J., Westendorp, R. G. J., and Blauw, G. J. (2006). In a population-based prospective study, no association between high blood pressure and mortality after age 85 years. *Journal of Hypertension*, 24(2):287–292.

Van Buuren, S. (2007a). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3):219–242.

Van Buuren, S. (2007b). Worm plot to diagnose fit in quantile regression. *Statistical Modelling*, 7(4):363–376.

Van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology*, 6(1):31–36.

Van Buuren, S. (2011). Multiple imputation of multilevel data. In Hox, J. and Roberts, J., editors, *The Handbook of Advanced Multilevel Analysis*, chapter 10, pages 173–196. Routledge, Milton Park, UK.

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12):1049–1064.

Van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2003). Assessing comparability of dressing disability in different countries by response conversion. *European Journal of Public Health*, 13(3 Suppl.):15–19.

Van Buuren, S., Eyres, S., Tennant, A., and Hopman-Rock, M. (2005). Improving comparability of existing data by response conversion. *Journal of Official Statistics*, 21(1):53–72.

Van Buuren, S. and Fredriks, A. M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20(8):1259–1277.

Van Buuren, S. and Groothuis-Oudshoorn, C. G. M. (1999). Flexible multivariate imputation by MICE. Technical Report PG/VGZ/99.054, TNO Prevention and Health, Leiden.

Van Buuren, S. and Groothuis-Oudshoorn, C. G. M. (2000). Multivariate imputation by chained equations: MICE V1.0 user's manual. Technical Report PG/VGZ/00.038, TNO Prevention and Health, Leiden.

Van Buuren, S. and Groothuis-Oudshoorn, C. G. M. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Van Buuren, S. and Ooms, J. C. L. (2009). Stage line diagram: An age-conditional reference diagram for tracking development. *Statistics in Medicine*, 28(11):1569–1579.

Van Buuren, S. and Tennant, A., editors (2004). *Response Conversion for the Health Monitoring Program*, volume 04 145. TNO Quality of Life, Leiden.

Van Buuren, S. and Van Rijckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency. *Psychometrika*, 57(4):567–580.

Van Buuren, S., Van Rijckevorsel, J. L. A., and Rubin, D. B. (1993). Multiple imputation by splines. In *Bulletin of the International Statistical Institute*, volume II (CP), pages 503–504.

Van Deth, J., editor (1998). *Comparative politics. The problem of equivalence.* Routledge, London.

Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2007). Multiple imputation for item scores when test data are factorially complex. *British Journal of Mathematical and Statistical Psychology*, 60(2):315–337.

Van Praag, B. M. S., Dijkstra, T. K., and Van Velzen, J. (1985). Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem. *Psychometrika*, 50(1):25–36.

Van Wouwe, J. P., Lanting, C. I., Van Dommelen, P., Treffers, P. E., and Van Buuren, S. (2009). Breastfeeding duration related to practised contraception in the Netherlands. *Acta Paediatrica*, 98(1):86–90.

Vateekul, P. and Sarinnapakorn, K. (2009). Tree-based approach to missing data imputation. In *2009 IEEE International Conference on Data Mining Workshops*, pages 70–75. IEEE Computer Society.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer-Verlag, New York, 4th edition.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Vergouw, D., Heymans, M. W., Peat, G. M., Kuijpers, T., Croft, P. R., De Vet, H. C. W., Van der Horst, H. E., and Van der Windt, D. A. W. M. (2010). The search for stable prognostic models in multiple imputed data sets. *BMC Medical Research Methodology*, 10:81.

Vergouwe, Y., Royston, P., Moons, K. G. M., and Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: A practical approach. *Journal of Clinical Epidemiology*, 63(2):205–214.

Verloove-Vanhorick, S. P., Verwey, R. A., Brand, R., Bennebroek Gravenhorst, J., Keirse, M. J. N. C., and Ruys, J. H. (1986). Neonatal mortality risk in relation to gestational age and birthweight. results of a national survey of preterm and very-low-birthweight infants in the Netherlands. *Lancet*, 1(8472):55–57.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1):369–397.

Viallefont, V., Raftery, A. E., and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine*, 20(21):3215–3230.

Visscher, T. L. S., Viet, A. L., Kroesbergen, H. T., and Seidell, J. C. (2006). Underreporting of BMI in adults and its effect on obesity prevalence estimations in the period 1998 to 2001. *Obesity*, 14(11):2054–2063.

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291.

Wallace, M. L., Anderson, S. J., and Mazumdar, S. (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in Medicine*, 29(29):3004–3016.

Walls, T. A. and Schafer, J. L., editors (2006). *Models for Intensive Longitudinal Data*. Oxford University Press, Oxford.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4):935–948.

Wang, Q. and Dinse, G. E. (2010). Linear regression analysis of survival data with missing censoring indicators. *Lifetime Data Analysis*, 17(2):256–279.

Wang, Y. J. and Kuo, K.-L. (2010). Compatibility of discrete conditional distributions with structural zeros. *Journal of Multivariate Analysis*, 101(1):191–199.

Wei, G. C. G. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 47(4):1297–1309.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, 29(28):2920–2931.

White, I. R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis*, 54(10):2267–2275.

White, I. R., Horton, N. J., Carpenter, J. R., and Pocock, S. J. (2011a). Strategy for intention to treat analysis in randomised trials with missing outcome data. *British Medical Journal*, 342:d40.

White, I. R. and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 28(15):1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011b). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399.

White, I. R. and Thompson, S. G. (2005). Adjusting for partially missing baseline measurements in randomized trials. *Statistics in Medicine*, 24(7):993–1007.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, 49:587–602.

Wood, A. M., White, I. R., and Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17):3227–3246.

Wood, A. M., White, I. R., and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1(4):368–376.

Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC, Boca Raton, FL.

Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61(2):498–506.

Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empirical Journal of Experimental Agriculture*, 1(2):129–142.

Yu, L.-M., Burton, A., and Rivero-Arias, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, 16(3):243–258.

Yucel, R. M., He, Y., and Zaslavsky, A. M. (2008). Using calibration to improve rounding in imputation. *The American Statistician*, 62(2):125–129.

Yucel, R. M. and Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association*, 100(472):1123–1132.

Zhang, S. (2011). Shell-neighbor method and its application in missing data imputation. *Applied Intelligence*, 35(1):123–133.

Zhang, W., Zhang, Y., Chaloner, K., and Stapleton, J. T. (2009). Imputation methods for doubly censored HIV data. *Journal of Statistical Computation and Simulation*, 79(10):1245–1257.

Zhao, E. and Yucel, R. M. (2009). Performance of sequential imputation method in multilevel applications. In *ASA 2009 Proceedings of the Survey Research Methods Section*, pages 2800–2810, Alexandria, VA.

This page intentionally left blank

Chapman & Hall/CRC Interdisciplinary Statistics Series

Missing data form a problem in every scientific discipline, yet the techniques required to handle them are complicated and often lacking. One of the great ideas in statistical science—multiple imputation—fills gaps in the data with plausible values, the uncertainty of which is coded in the data itself. It also solves other problems, many of which are missing data problems in disguise.

Flexible Imputation of Missing Data is supported by many examples using real data taken from the author's vast experience of collaborative research, and presents a practical guide for handling missing data under the framework of multiple imputation. Furthermore, detailed guidance of implementation in R using the author's package MICE is included throughout the book.

Assuming familiarity with basic statistical concepts and multivariate methods, **Flexible Imputation of Missing Data** is intended for two audiences:

- (Bio)statisticians, epidemiologists, and methodologists in the social and health sciences
- Substantive researchers who do not call themselves statisticians, but who possess the necessary skills to understand the principles and to follow the recipes

This graduate-tested book avoids mathematical and technical details as much as possible: formulas are accompanied by a verbal statement that explains the formula in layperson terms. Readers less concerned with the theoretical underpinnings will be able to pick up the general idea, and technical material is available for those who desire deeper understanding. The analyses can be replicated in R using a dedicated package developed by the author.


CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K13103
ISBN: 978-1-4398-6824-9
90000
9 781439 868249