JAMES R. CARPENTER
MICHAEL G. KENWARD

Multiple Imputation and its Application

WILEY

STATISTICS IN PRACTICE

Multiple Imputation and its Application

Statistics in Practice

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

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

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

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

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

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

Multiple Imputation and its Application

James R. Carpenter

and

Michael G. Kenward

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

WILEY

A John Wiley & Sons, Ltd., Publication

This edition first published 2013

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

Library of Congress Cataloging-in-Publication Data

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

2012028821

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

ISBN: 978-0-470-74052-1

Cover photograph courtesy of Harvey Goldstein

Set in 10/12pt Times by Laserwords Private Limited, Chennai, India

Contents

Preface

No study of any complexity manages to collect all the intended data. Analysis of the resulting partially collected data must therefore address the issues raised by the missing data. Unfortunately, the inferential consequences of missing data are not simply restricted to the proportion of missing observations. Instead, the interplay between the substantive questions and the reasons for the missing data is crucial. Thus, there is no simple, universal, solution.

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

Given a set of assumptions about the reasons data are missing, there are a number of statistical methods for carrying out the analysis. These include the EM algorithm, inverse probability weighting, a full Bayesian analysis and, depending on the setting, a direct application of maximum likelihood. These methods, and those derived from them, each have their own advantages in particular settings. Nevertheless, we argue that none shares the practical utility, broad applicability and relative simplicity of Rubin's Multiple Imputation (MI).

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

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

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

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

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

Data acknowledgements

We are grateful to the following:

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

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

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

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

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

In Chapters 1, 5, 8, 10 and 11 we have analysed data from the Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 First Edition, Colchester, Essex, published by and freely available from the UK Data Archive, Study Number SN 5765. We thank Vernon Gayle for introducing us to these data.

In Chapter 6 we have analysed data from the Alzheimer's Disease Neuro-imaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or in the writing of this book. A complete listing of ADNI investigators can be found at: `http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`.

The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and nonprofit organisations, as a $60 million, five-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the

progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

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

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

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

Acknowledgements

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

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

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

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

Thanks to Tim Collier for the anecdote in §1.3.

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

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

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

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

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

Glossary

Indices and symbols

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

Matrices

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

Abbreviations

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

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

Probability distributions

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

PART I

FOUNDATIONS

1

Introduction

Collecting, analysing and drawing inferences from data are central to research in the medical and social sciences. Unfortunately, for any number of reasons, it is rarely possible to collect all the intended data. The ubiquity of missing data, and the problems this poses for both analysis and inference, has spawned a substantial statistical literature dating from 1950s. At that time, when statistical computing was in its infancy, many analyses were only feasible because of the carefully planned balance in the dataset (for example, the same number of observations on each unit). Missing data meant the available data for analysis were unbalanced, thus complicating the planned analysis and in some instances rendering it unfeasible. Early work on the problem was therefore largely computational (e.g. Healy and Westmacott, 1956; Afifi and Elashoff, 1966; Orchard and Woodbury, 1972; Dempster *et al*., 1977).

The wider question of the consequences of nontrivial proportions of missing data for inference was neglected until a seminal paper by Rubin (1976). This set out a typology for assumptions about the reasons for missing data, and sketched their implications for analysis and inference. It marked the beginning of a broad stream of research about the analysis of partially observed data. The literature is now huge, and continues to grow, both as methods are developed for large and complex data structures, and as increasing computer power and suitable software enable researchers to apply these methods.

For a broad overview of the literature, a good place to start is one of the recent excellent textbooks. Little and Rubin (2002) write for applied statisticians. They give a good overview of likelihood methods, and give an introduction to multiple imputation. Allison (2002) presents a less technical overview. Schafer (1997) is more algorithmic, focusing on the EM algorithm and imputation using the multivatiate normal and general location model. Molenberghs and Kenward (2007)

Multiple Imputation and its Application, First Edition. James R. Carpenter and Michael G. Kenward.

focus on clinical studies, while Daniels and Hogan (2008) focus on longitudinal studies with a Bayesian emphasis.

The above books concentrate on parametric approaches. However, there is also a growing literature based around using inverse probability weighting, in the spirit of Horvitz and Thompson (1952), and associated doubly robust methods. In particular, we refer to the work of Robins and colleagues (e.g. Robins *et al*., 1995; Scharfstein *et al*., 1999). Vansteelandt *et al*. (2009) give an accessible introduction to these developments. A comparison with multiple imputation in a simple setting is given by Carpenter *et al*. (2006). The pros and cons are debated in Kang and Schafer (2007) and the theory is brought together by Tsiatis (2006).

This book is concerned with a particular statistical method for analysing and drawing inferences from incomplete data, called *Multiple Imputation (MI)*. Initially proposed by Rubin (1987) in the context of surveys, increasing awareness among researchers about the possible effects of missing data (e.g. Klebanoff and Cole, 2008) has led to an upsurge of interest (e.g. Sterne *et al*., 2009; Kenward and Carpenter, 2007; Schafer, 1999a; Rubin, 1996).

Multiple imputation (MI) is attractive because it is both practical and widely applicable. Recently developed statistical software (see, for example, issue 45 of the *Journal of Statistical Software*) has placed it within the reach of most researchers in the medical and social sciences, whether or not they have undertaken advanced training in statistics. However, the increasing use of MI in a range of settings beyond that originally envisaged has led to a bewildering proliferation of algorithms and software. Further, the implication of the underlying assumptions in the context of the data at hand is often unclear.

We are writing for researchers in the medical and social sciences with the aim of clarifying the issues raised by missing data, outlining the rationale for MI, explaining the motivation and relationship between the various imputation algorithms, and describing and illustrating its application to increasingly complex data structures.

Central to the analysis of partially observed data is an understanding of why the data are missing and the implications of this for the analysis. This is the focus of the remainder of this chapter. Introducing some of the examples that run through the book, we show how Rubin's typology (Rubin, 1976) provides the foundational framework for understanding the implications of missing data.

1.1 Reasons for missing data

In this section we consider possible reasons for missing data, illustrate these with examples, and draw some preliminary implications for inference. We use the word 'possible' advisedly, since with partially observed data we can rarely be sure of the mechanism giving rise to missing data. Instead, a range of possible mechanisms are consistent with the observed data. In practice, we therefore wish to analyse the data under different mechanisms, to establish the robustness of our inference in the face of uncertainty about the missingness mechanism.

All datasets consist of a series of *units* each of which provides information on a series of *items*. For example, in a cross-sectional questionnaire survey, the units would be individuals and the items their answers to the questions. In a household survey, the units would be households, and the items information about the household and members of the household. In longitudinal studies, units would typically be individuals while items would be longitudinal data from those individuals. In this book, units therefore correspond to the highest level in multilevel (i.e., hierarchical) data, and unless stated otherwise data from different units are statistically independent.

Within this framework, it is useful to distinguish between units where all the information is missing, termed *unit nonresponse* and units who contribute partial information, termed *item nonresponse*. The statistical issues are the same in both cases, and both can in principle be handled by MI. However, the main focus of this book is the latter.

Example 1.1 Mandarin tableau

Figure 1.1, which is also shown on the cover, shows part of the frontage of a senior mandarin's house in the New Territories, Hong Kong. We suppose interest focuses on characteristics of the figurines, for example their number, height, facial characteristics and dress. Unit nonresponse then corresponds to missing figurines, and item nonresponse to damaged – hence partially observed – figurines. □

Figure 1.1 Detail from a senior mandarin's house front in New Territories, Hong Kong. Photograph by H. Goldstein.

1.2 Examples

We now introduce two key examples, which we return to throughout the book.

Example 1.2 Youth Cohort Study (YCS)

The Youth Cohort Study of England and Wales (YCS) is an ongoing UK government funded representative survey of pupils in England and Wales at school-leaving age (School year 11, age 16–17) (UK Data Archive, 2007). Each year that a new cohort is surveyed, detailed information is collected on each young person's experience of education and their qualifications as well as information on employment and training. A limited amount of information is collected on their personal characteristics, family, home circumstances, and aspirations.

Over the life-cycle of the YCS, different organisations have had responsibility for the structure and timings of data collection. Unfortunately, the documentation of older cohorts is poor. Croxford *et al*. (2007) have recently deposited a harmonised dataset that comprises YCS cohorts from 1984 to 2002 (UK Data Archive Study Number 5765). We consider data from pupils attending comprehensive schools from five YCS cohorts; these pupils reached the end of Year 11 in 1990, 1993, 1995, 1997 and 1999.

We explore relationships between Year 11 educational attainment (the General Certificate of Secondary Education) and key measures of social stratification. The units are pupils and the items are measurements on these pupils, and a nontrivial number of items are partially observed. □

Example 1.3 Randomised controlled trial of patients with chronic asthma

We consider data from a 5-arm asthma clinical trial to assess the efficacy and safety of budesonide, a second-generation glucocorticosteroid, on patients with chronic asthma. 473 patients with chronic asthma were enrolled in the 12-week randomised, double-blind, multi-centre parallel-group trial, which compared the effect of a daily dose of 200, 400, 800 or 1600 mcg of budesonide with placebo.

Key outcomes of clinical interest include patients' peak expiratory flow rate (their maximum speed of expiration in litres/minute) and their Forced Expiratory Volume, FEV_1, (the volume of air, in litres, the patient with fully inflated lungs can breathe out in one second). In summary, the trial found a statistically significant dose-response effect for the mean change from baseline over the study for both morning peak expiratory flow, evening peak expiratory flow and FEV_1, at the 5% level.

Budesonide treated patients also showed reduced asthma symptoms and bronchodilator use compared with placebo, while there were no clinically significant differences in treatment related adverse experiences between the treatment groups. Further details about the conduct of the trial, its conclusions and the variables collected can be found elsewhere (Busse *et al*., 1998). Here, we focus on FEV_1 and confine our attention to the placebo and lowest active

dose arms. FEV_1 was collected at baseline, then 2, 4, 8 and 12 weeks after randomisation. The intention was to compare FEV_1 across treatment arms at 12 weeks. However, excluding 3 patients whose participation in the study was intermittent, only 37 out of 90 patients in the placebo arm, and 71 out of 90 patients in the lowest active dose arm, still remained in the trial at twelve weeks. □

1.3 Patterns of missing data

It is very important to investigate the patterns of missing data before embarking on a formal analysis. This can throw up vital information that might otherwise be overlooked, and may even allow the missing data to be traced. For example, when analysing the new wave of a longitudinal survey, a colleague's careful examination of missing data patterns established that many of the missing questionnaires could be traced to a set of cardboard boxes. These turned out to have been left behind in a move. They were recovered and the data entered.

Most statistical software now has tools for describing the pattern of missing data. Key questions concern the extent and patterns of missing values, and whether the pattern is *monotone* (as described in the next paragraph), as if it is, this can considerably speed up and simplify the analysis.

Missing data in a set of p variables are said to follow a *monotone missingness pattern* if the variables can be re-ordered such that, for every unit i and variable j,

1. if unit i is observed on variable j, where $j = 2, \ldots, p$, it is observed on all variables $j' < j$, and
2. if unit i is missing on variable j, where $j = 2, \ldots, p$, it is missing on all variables $j' > j$.

A natural setting for the occurrence of monotone missing data is a longitudinal study, where units are observed either until they are lost to follow-up, or the study concludes. A monotone pattern is thus inconsistent with interim missing data, where units are observed for a period, missing for the subsequent period, but then observed. Questionnaires may also give rise to monotone missing data patterns when individuals systematically answer each question in turn from the beginning till they either stop or complete the questionnaire. In other settings it may be possible to re-order items to achieve a monotone pattern.

Example 1.2 Youth Cohort Study *(ctd)*

Table 1.1 shows the covariates we consider from the YCS. There are no missing data in the variables *cohort* and *boy*. The missingness pattern for GCSE score and the remaining two variables is shown in Table 1.2. In this example it is not possible to re-order the variables (items) to obtain a monotone pattern, due for example, to pattern 3 (N = 697). □

Table 1.1 YCS variables for exploring the relationship between Year 11 attainment and social stratification.

Variable name	Description
cohort	year of data collection: 1990, 93, 95, 97, 99
boy	indicator variable for boys
occupation	parental occupation, categorised as managerial, intermediate or working
ethnicity	categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Table 1.2 Pattern of missing values in the YCS data.

Pattern	GCSE score	Occupation	Ethnicity	No.	% of total
1	✓	✓	✓	55145	87%
2	✓	·	✓	6821	11%
3	·	✓	✓	697	1%
4	✓	·	·	592	1%

Example 1.3 Asthma study *(ctd)*

Table 1.3 shows the withdrawal pattern for the placebo and lowest active dose arms (all the patients are receiving their randomised medication). We have removed three patients with unusual interim missing data from Table 1.3 and all our analyses. The remaining missingness pattern is monotone in both treatment arms. □

Table 1.3 Asthma study: withdrawal pattern by treatment arm.

Dropout pattern	Placebo arm						
	Mean FEV_1 (litres) measured at week					Number	Percent
	0	2	4	8	12		
1	✓	✓	✓	✓	✓	37	41
2	✓	✓	✓	✓	·	15	17
3	✓	✓	✓	·	·	22	24
4	✓	✓	·	·	·	16	18
	Lowest Active arm						
1	✓	✓	✓	✓	✓	71	79
2	✓	✓	✓	✓	·	8	9
3	✓	✓	✓	·	·	8	9
4	✓	✓	·	·	·	3	3

1.3.1 Consequences of missing data

Our focus is the practical implications of missing data for both parameter estimation and inference. Unfortunately, the two are often conflated, so that a computational method for parameter estimation when data are missing is said to have 'solved' or 'handled' the missing data issue. Since, with missing data, computational methods only lead to valid inference under specific assumptions, this attitude is likely to lead to misleading inferences.

In this context, it may be helpful to draw an analogy with the sampling process used to collect the data. If an analyst is presented with a spreadsheet containing columns of numerical data, they can analyse the data (calculate means of variables, regress variables on each other and so forth). However, they cannot draw any inferences unless they are told how and from whom the data were collected. This information is external to the numerical values of the variables.

We may think of the missing data mechanism as a second stage in the sampling process, but one that is not under our control. It acts on the data we intended to collect and leaves us with a partially observed dataset. Once again, the missing data mechanism cannot usually be definitively identified from the observed data, although the observed data may indicate plausible mechanisms (e.g. response may be negatively correlated with age). Thus we will need to make an assumption about the missingness mechanism in order to draw inference. The process of making this assumption is quite separate from the statistical methods we use for parameter estimation etc. Further, to the extent that the missing data mechanism cannot be definitively identified from the data, we will often wish to check the robustness of our inferences to a range of missingness mechanisms that are consistent with the observed data. The reason this book focuses on the statistical method of MI is that it provides a computationally feasible approach to the analysis for a wide range of problems under a range of missingness mechanisms.

We therefore begin with a typology for the mechanisms causing, or generating, the missing data. Later in this chapter we will see that consideration of these mechanisms in the context of the analysis at hand clarifies the assumptions under which a simple analysis, such as restriction to complete records, will be valid. It also clarifies when more sophisticated computational approaches such as MI will be valid and informs the way they are conducted. We stress again that the mechanism causing the missing data can rarely be definitively established. Thus we will often wish to explore the robustness of our inferences to a range of plausible missingness mechanisms – a process we call *sensitivity analysis*.

From a general standpoint, missing data may cause two problems: loss of efficiency and bias.

First, loss of efficiency, or information, is an inevitable consequence of missing data. Unfortunately, the extent of information loss is not directly linked to the proportion of incomplete records. Instead it is intrinsically linked to the analysis question. When crossing the road, the rear of the oncoming traffic is hidden from view – the data are missing. However, these missing data do not bear on the question at hand – will I make it across the road safely? While the proportion

of missing data about each oncoming vehicle is substantial, information loss is negligible. Conversely, when estimating the prevalence of a rare disease, a small proportion of missing observations could have a disproportionate impact on the resulting estimate.

Faced with an incomplete dataset, most software automatically restricts analysis to complete records. As we illustrate below, the consequence of this for loss of information is not always easy to predict. Nevertheless, in many settings it will be important to include the information from partially complete records. Not least of the reasons for this is the time and money it has taken to collect even the partially complete records. Under certain assumptions about the missingness mechanism, we shall see that MI provides a natural way to do this.

Second, and perhaps more fundamentally, the subset of complete records may not be representative of the population under study. Restricting analysis to complete records may then lead to biased inference. The extent of such bias depends on the statistical behaviour of the missing data. A formal framework to describe this behaviour is thus fundamental. Such a framework was first elucidated in a seminal paper by Rubin (1976). To describe this, we need some definitions.

1.4 Inferential framework and notation

For clarity we take a frequentist approach to inference. This is not essential or necessarily desirable; indeed we will see that MI is essentially a Bayesian method, with good frequentist properties. Often, as Chapter 2 shows, formally showing these frequentist properties is most difficult theoretically.

We suppose we have a sample of n units, which will often be individuals, from a population that for practical inferential purposes can be considered infinite. Let $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,p})^T$ denote the p variables we intended to collect from the i^{th} unit, $i = 1 \ldots, n$. We wish to use these data to make inferences about a set of p population parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$.

For each unit $i = 1, \ldots, n$ let $\mathbf{Y}_{i,O}$ denote the subset of p variables that are observed, and $\mathbf{Y}_{i,M}$ denote the subset that are missing. Thus, for different individuals, $\mathbf{Y}_{i,O}$ and $\mathbf{Y}_{i,M}$ may well be different subsets of the p variables. If no data are missing, $\mathbf{Y}_{i,M}$ will be empty.

Next, again for each individual $i = 1, \ldots, n$ and variable $j = 1, \ldots, p$, let $R_{i,j} = 1$ if $Y_{i,j}$ is observed and $R_{i,j} = 0$ if $Y_{i,j}$ is missing. Let $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,p})^T$. Consistent with the definition of monotone missingness patterns on p. 10, the pattern is monotone if the p variables can be re-ordered so that for each unit i,

$$R_{i,j} = 0 \implies R_{i,j'} = 0 \text{ for } j' = j + 1, \ldots, p. \tag{1.1}$$

The missing value mechanism is then formally defined as

$$\Pr(\mathbf{R}_i | \mathbf{Y}_i), \tag{1.2}$$