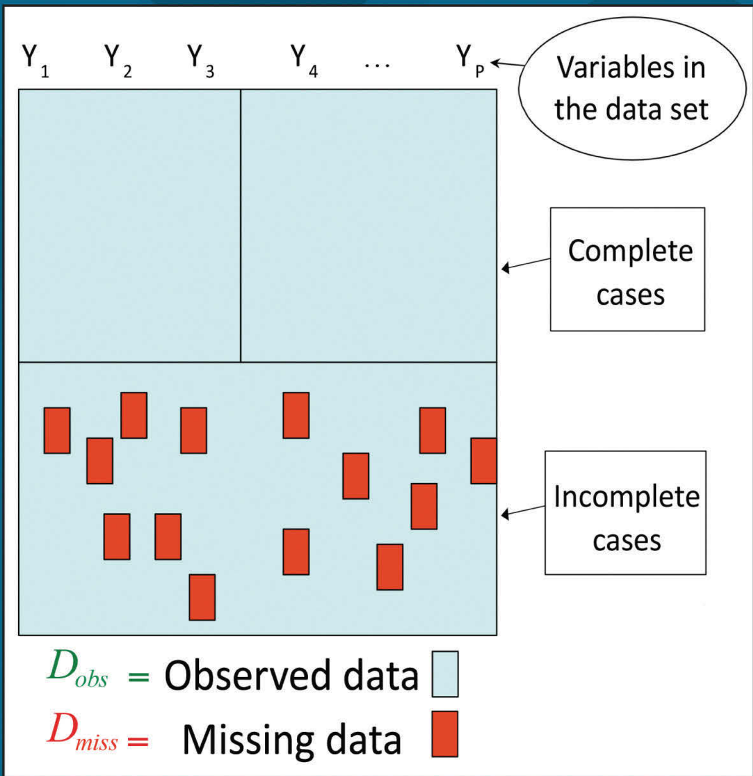
Chapman & Hall/CRC
Interdisciplinary Statistics Series

Missing Data Analysis in Practice



Trivellore Raghunathan

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Missing Data Analysis
in Practice

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

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

Published titles

AGE-PERIOD-COHORT ANALYSIS: NEW MODELS, METHODS, AND EMPIRICAL APPLICATIONS Y. Yang and K. C. Land

ANALYSIS OF CAPTURE-RECAPTURE DATA R. S. McCrea and B. J.T. Morgan

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES S. Lele and J. Richtsmeier

ASTROSTATISTICS G. Babu and E. Feigelson

BAYESIAN ANALYSIS FOR POPULATION ECOLOGY R. King, B. J.T. Morgan, O. Gimenez, and S. P. Brooks

BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY, SECOND EDITION A. B. Lawson

BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY S. Patterson and B. Jones

CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION S. Green, J. Benedetti, A. Smith, and J. Crowley

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

INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES M. Waterman

MARKOV CHAIN MONTE CARLO IN PRACTICE W. Gilks, S. Richardson, and D. Spiegelhalter

MEASUREMENT ERROR ANDMISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS P. Gustafson

MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS J. P. Buonaccorsi

MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS J. P. Buonaccorsi

Published titles

MENDELIAN RANDOMIZATION: METHODS FOR USING GENETIC VARIANTS IN CAUSAL ESTIMATION S.Burgess and S.G. Thompson

META-ANALYSIS OF BINARY DATA USINGPROFILE LIKELIHOOD D. Böhning, R. Kuhnert, and S. Rattanasiri

MISSING DATA ANALYSIS IN PRACTICE T. Raghunathan

POWER ANALYSIS OF TRIALS WITH MULTILEVEL DATA M. Moerbeek and S. Teerenstra

SPATIAL POINT PATTERNS: METHODOLOGY AND APPLICATIONS WITH R A. Baddeley, E Rubak, and R. Turner

STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA T. Speed

STATISTICAL ANALYSIS OF QUESTIONNAIRES: A UNIFIED APPROACH BASED ON R AND STATA F. Bartolucci, S. Bacci, and M. Gnaldi

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

Chapman & Hall/CRC
Interdisciplinary Statistics Series

Missing Data Analysis in Practice

Trivellore Raghunathan
University of Michigan
Ann Arbor, Michigan, USA

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

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

Version Date: 20150915

International Standard Book Number-13: 978-1-4822-1193-1 (eBook - PDF)

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

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

To my Teachers
and my Students.

Contents

List of Tables

List of Figures

Preface

Missing data problems are ubiquitous and as old as data analysis itself. Researchers have been dealing with missing data in various ways and in a somewhat *ad hoc* manner. While some *ad hoc* procedures have taken roots, principled approaches such as maximum likelihood were developed in the middle of the last century. Lack of computational power and to some extent sheer inertia against any change led to a confusing landscape. Many principled methods were not implemented in familiar software packages. The last quarter of the last century witnessed massive changes in the computational landscape and large scale applications of the statistical methodology. This allowed methodological development for handling incomplete data, Bayesian methods, replication or resampling methods, complex modeling and their implementation to a desktop, to a laptop and even to a palm of one hand!

Now researchers are like children in a toy store! Too many methods, too many implementations and still a confusing landscape. Every missing data method makes assumptions. This is not bad and is just inherent in making inferences with incomplete data. The design of a study may be under the control of the investigator, but responses from subjects in the study or their cooperation to provide responses are not. In fact, it can be argued that assumptions are inherent in any statistical inferential activity as a projection from a sample to the population or the phenomenon. It is important, therefore, to know those assumptions and to discriminatively assess the suitability of assumptions in the specific context.

This book describes several easy-to-implement approaches, discusses the underlying assumptions, provides some practical means for assessing these assumptions and then suggests implementation. Numerous approximations are used as no practical problem comes in a nice theoretically elegant mold. The theory is described using heuristics rather than rigor. Actual and simulated data sets are used to illustrate important concepts.

This book uses ideas from both Frequentist and Bayesian perspectives but has a definite Bayesian flavor. A statistical practitioner is like a "handyman," a person interested in solving the practical problems and no good tool, frequentist or Bayesian, is wasteful. All distributional assumptions are stated so that they can be interpreted from both perspectives. The danger with this attitude is that both camps may be unhappy with the book!

Many books on incomplete data have been published over the last few years. Why another book? Some books are a bit too theoretical and suitable for students in statistics or biostatistics programs. Some books are geared towards one approach or the other and/or one software or the other. Some are too general and lack specifics on how to apply. Over the last 25 years, I have taught several one-day to two semester long courses for a variety of audiences with a range of quantitative backgrounds. I have attempted to write this book for that kind of audience as if they are all in my lecture. I have described the methods in simple terms and in technical terms. This book attempts to provide heuristic reasoning of the assumptions, theoretical understanding and practical implementation. This book represents a collective experience of research, teaching and consulting.

A two semester course in statistics including regression analysis should suffice to understand most of the material. A course in Bayesian analysis is a big plus for a practitioner as it can expand the knowledge base for tackling many practical problems. A healthy Frequentist and Bayes concoction are a great way to start the statistical practice.

I have co-taught several two-day short courses with Rod Little and have developed Power-Point slides. I have greatly benefited by this collaboration. The first four chapters of the book mostly follow the presentation in the slides. The maximum likelihood (ML) approach is discussed in brief. Of course, Little and Rubin (2002) is the best book for ML! At Michigan, Rod has been a wonderful mentor, a dear friend and a superb colleague.

Though I received solid training in mathematical statistics at the Institute of Science, Nagpur in India, and a thorough knowledge of applications at Miami University in Oxford, Ohio, my statistical being was shaped at Harvard University by Don Rubin and Art Dempster. Fred Mosteller and John Tukey (who provided great inspiration and emphasized the importance of "listening to data" during his regular summer visits to Harvard) greatly influenced me. Being a teaching assistant for Fred Mosteller provided the unique perspective on teaching, nay, learning through and with the students. To this mix add some of the smartest people, Nat Schenker, Xiao-li Meng, Andy Gelman, Alan Zaslavsky, Emery Brown, Joe Schafer, Tom Belin and many others as colleagues at the graduate school, academic brothers and the continuing relationship beyond (Am I describing an academic heaven?). I owe my deep gratitude to these and many more. A special thanks to Nat Schenker who has been a great collaborator.

I thank Dawn Reed for compiling the references and other help with manuscript preparations. I am thankful to all the students in my classes who over the years brought their missing data problems to my attention, helped me understand the practical issues and shaped this book through their comments and through challenging some assertions. These discussions led to several simulation studies for the entire class. Many such simulation studies are discussed in the book.

The first seven chapters deal with the traditional missing data problem. However, many statistical problems with no missing values can be addressed using the missing data framework. Some of these applications are discussed in Chapter 8 and thus extending the application of methods discussed in this book. Missing data methods development and its adaptation to practical problems is, therefore, a great area of research and application. Hope that material presented in the book is useful!

Trivellore Raghunathan ("Raghu")
Ann Arbor, Michigan

1

Basic Concepts

1.1 Introduction

Data collection and analysis forms the backbone of all empirical research and almost every data analysis involves variables with some missing values (which will be defined later). The missing values may arise due to unit nonresponse where a sampled subject refuses to provide any values for the variables of interest, or due to item nonresponse, where a sampled subject provides information only for some variables.

The complete-case or available-subjects is the common approach that restricts the analysis to subjects without missing values in the relevant variables. This approach, though convenient, can result in biased estimates of the parameters or population quantities because the included and excluded subjects from the analysis may differ systematically. Even if the included subjects are a random subset of the sampled subjects, the sampling error increases due to the reduced sample size.

Many *ad hoc* and naive methods are also used in practice. For example, in a multiple regression analysis with missing values in one categorical covariate, subjects with missing values are treated as a separate category when creating the dummy variables. This analysis uses all of the subjects but may seriously bias the regression coefficients for the other covariates. Another naive method involves substituting a fixed value, such as the mean, median or the mode based on respondents for all subjects, with missing values. This strategy creates artificial "peaks" in the imputed (or the completed) data set, resulting in bias in any analysis involving statistics measuring spread or dispersion, such as regression analysis.

Despite several studies demonstrating bias through theoretical and simulation investigations in using complete-case and *ad hoc* methods, they continue to be used. There may be special circumstances or assumptions under which