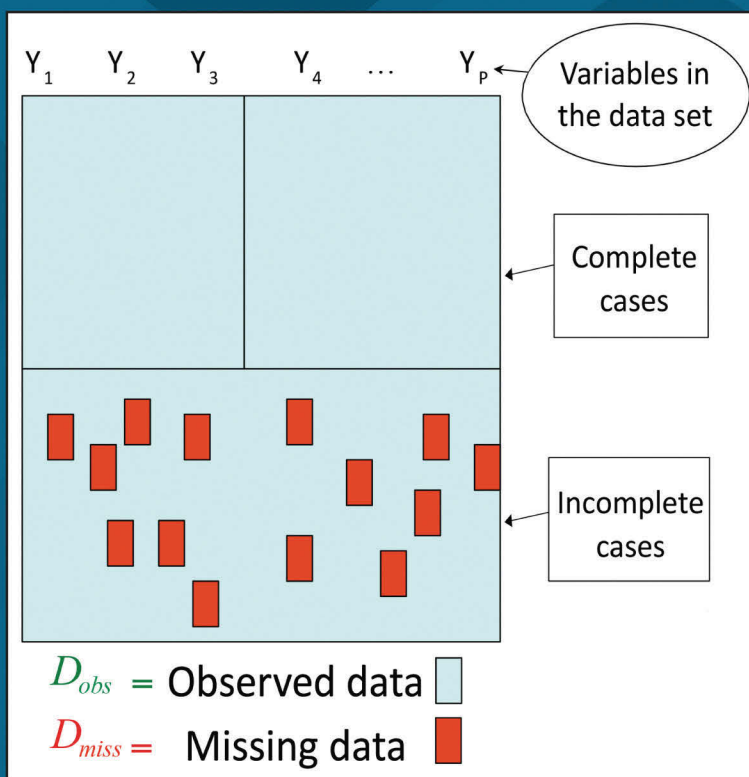


Missing Data Analysis in Practice



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, LONGITUDI-
NAL, 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 AND MISCLASSIFICATION IN STATISTICS AND EPIDE-
MIOLOGY: 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 USING PROFILE 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

© 2016 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: 20150915

International Standard Book Number-13: 978-1-4822-1193-1 (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>

*To my Teachers
and my Students.*

Contents

List of Tables	xiii
List of Figures	xv
Preface	xvii
1 Basic Concepts	1
1.1 Introduction	1
1.2 Definition of Missing Values	2
1.3 Missing Data Pattern	3
1.4 Missing Data Mechanism	4
1.5 Problems with Complete-Case Analysis	7
1.6 Analysis Approaches	9
1.7 Basic Statistical Concepts	13
1.8 A Chuckle or Two	19
1.9 Bibliographic Note	21
1.10 Exercises	23
2 Weighting Methods	27
2.1 Motivation	27
2.2 Adjustment Cell Method	29
2.3 Response Propensity Model	29
2.4 Example	32
2.5 Impact of Weights on Population Mean Estimates	37
2.6 Post-Stratification	39
2.6.1 Post-Stratification Weights	39
2.6.2 Raking	40
2.6.3 Post-stratified Estimator	42
2.7 Survey Weights	44
2.8 Alternative to Weighted Analysis	45

2.9	Inverse Probability Weighting	47
2.10	Bibliographic Note	47
2.11	Exercises	49
3	Imputation	51
3.1	Generation of Plausible Values	53
3.2	Hot Deck Imputation	55
3.2.1	Connection with Weighting	57
3.2.2	Bayesian Modification	58
3.3	Model Based Imputation	59
3.4	Example	63
3.5	Sequential Regression Imputation	67
3.5.1	Details	69
3.5.2	Handling Restrictions	71
3.5.3	Model Fitting Issues	73
3.6	Bibliographic Note	75
3.7	Exercises	76
4	Multiple Imputation	77
4.1	Introduction	77
4.2	Basic Combining Rule	77
4.3	Multivariate Hypothesis Testing	79
4.4	Combining Test Statistics	80
4.5	Basic Theory of Multiple Imputation	82
4.6	Extended Combining Rules	83
4.6.1	Transformation	84
4.6.2	Nonnormal Approximation	85
4.7	Some Practical Issues	86
4.7.1	Number of Imputations	86
4.7.2	Diagnostics	87
4.7.3	To Impute or Not to Impute	88
4.8	Revisiting Examples	89
4.8.1	Data in Table 1.1	89
4.8.2	Case-Control Study	90
4.9	Example: St. Louis Risk Research Project	91
4.10	Bibliographic Note	95
4.11	Exercises	95

5	Regression Analysis	99
5.1	General Observations	99
5.1.1	Imputation Issues	99
5.2	Revisiting St. Louis Risk Research Example	103
5.3	Analysis of Variance	105
5.3.1	Complete Data Analysis	106
5.3.1.1	Partitioning of Sum of Squares	106
5.3.1.2	Regression Formulation	108
5.3.2	ANOVA with Missing Values	108
5.3.2.1	Combining Sums of Squares	109
5.3.2.2	Regression Formulation with Missing Values	110
5.3.3	Example	110
5.3.4	Extensions	112
5.4	Survival Analysis Example	113
5.5	Bibliographic Note	117
5.6	Exercises	117
6	Longitudinal Analysis with Missing Values	121
6.1	Introduction	121
6.2	Imputation Model Assumption	124
6.2.1	Completed as Randomized	126
6.2.2	Completed as Control	128
6.2.3	Completed as Stable	129
6.3	Example	130
6.3.1	Completed as Randomized: Maximum Likelihood Analysis	130
6.3.2	Multiple Imputation: Completed as Randomized	133
6.3.3	Multiple Imputation: Completed as Control	135
6.4	Practical Issues	135
6.5	Weighting Methods	136
6.6	Binary Example	139
6.7	Bibliographic Note	142
6.8	Exercises	143
7	Nonignorable Missing Data Mechanisms	145
7.1	Modeling Framework	145
7.2	EM-Algorithm	146

7.3	Inference under Selection Model	148
7.4	Inference under Mixture Model	151
7.5	Example	151
7.6	Practical Considerations	152
7.7	Bibliographic Note	153
7.8	Exercises	154
8	Other Applications	155
8.1	Measurement Error	155
8.2	Combining Information from Multiple Data Sources	159
8.3	Bayesian Inference from Finite Population	160
8.4	Causal Inference	163
8.5	Disclosure Limitation	165
8.6	Bibliographic Note	169
8.7	Problems	170
9	Other Topics	175
9.1	Uncongeniality and Multiple Imputation	175
9.2	Multiple Imputation for Complex Surveys	177
9.3	Missing Values by Design	179
9.4	Replication Method for Variance Estimation	180
9.5	Final Thoughts	182
9.6	Bibliographic Note	183
9.7	Exercises	184
	Bibliography	187

--- *List of Tables* ---

1.1	Descriptive statistics based on simulated before and after deletion data sets	8
2.1	Mean (SD) or proportion for the six individual level variables by response status	33
2.2	Mean (SD) of ten housing or block level variables by response status	34
2.3	Nonresponse adjustment weights based propensity score stratification	36
2.4	Unweighted and weighted frequency distributions (in %) and their standard errors of self-reported health	37
2.5	Construction of post-stratification weights	40
2.6	Post-stratification example with raked weights	40
2.7	Sample proportions reported not having any health insurance	42
3.1	Mean (SD) of red-cell membrane and dietary intake of omega-3 fatty acids	65
3.2	Summary statistics of the observed and imputed logarithm of the red-cell membrane omega-3 fatty acids by case-control status	67
4.1	Maternal smoking and child wheeze status from the six city study	81
4.2	Cell specific percentages (SE) [FMI%] for the data from the six city study	82
4.3	Estimates and their standard errors for logistic regression example using the simulated data in Table 1.1	90
4.4	Estimated regression coefficient, the standard error (SE) and the fraction of missing information (FMI) for the case-control study example based on $M = 100$ imputations	90

4.5	Estimates, standard errors and the fraction of missing information for estimating the mean	91
4.6	Multiple imputation mean, standard error and the degrees of freedom for three composite variables R , V and S for three groups	93
4.7	Fraction of missing observations and missing information: St. Louis risk research study	94
5.1	Multiply imputed random effects model analysis of reading and verbal scores in the St. Louis risk research study	104
5.2	Multiply imputed generalized estimating equation analysis of symptoms in the St. Louis risk research study	105
5.3	Results from complete-case and two multiple imputation analysis of variance methods	112
5.4	Description of the variables and the number of missing values. Censored observations are treated as missing values	114
5.5	Results from the multiple imputation analysis of PBC data	116
6.1	Mean and standard deviation of visual analog scale score by treatment group and days of treatment	131
6.2	Maximum likelihood estimate of the parameters	132
6.3	Covariate adjusted multiple imputation mean difference between the Bup-Nx and clonidine groups for each day of the study, its standard error and the fraction of missing information	134
6.4	Summary statistics based on the observed ACL data: (number missing) and percent with cognitive impairment. Only non-missing cases in Wave 1 included. No age restrictions applied.	140
6.5	Comparison of intercepts and slopes between African-Americans and whites in each socio-economic status group. Baseline age 50 years or younger.	142
7.1	Estimated treatment effect (SE) on Day 14 visual analog score (VAS) adjusted for age, gender, race and baseline VAS	152

List of Figures

1.1	Patterns of missing data	4
1.2	Results from a simulation study of logistic regression analysis with missing covariates	9
2.1	Histogram of the non response adjustment weights	36
2.2	Box plot of the response propensity weights by self-rated health status	38
3.1	Scatter plot of red-cell and dietary values of omega-3 fatty acids on log-scale	64
3.2	Comparison of observed and imputed values	66
4.1	Scatter plots of variables in St. Louis risk research project . .	93
5.2	Scatter plots comparing estimates with and without expendi- tures in the imputation model: CEX simulation study	103
5.3	Diagnostic scatter plots to check the distribution of the ob- served and imputed values in the PBC analysis	115
6.1	Longitudinal study comparing Bup-Nx and clonidine for detox- fication	131
6.2	Longitudinal study comparing Bup-Nx and clonidine for detox- fication	133
6.3	Estimated daily mean difference between Bup-Nx and clonidine groups	134
6.4	Least squares regression lines relating the proportion of subjects with impairment as a function of time (in decades) for the 8 SES-race group. Baseline age 50 years or younger.	141
8.1	Measurement error study scenarios	156

8.2 Scatter plot total cultivated area in 1931 and area under wheat
in 1936, both on the square root scale 162

8.3 Data structure in the causal inference 164

8.4 Schematic display of the process for creating fully synthetic
data sets 167

8.5 Variants of synthetic data sets to limit statistical disclosure . 169

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

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