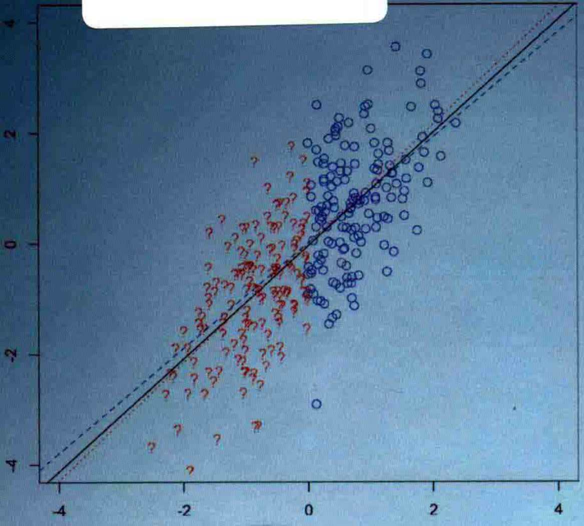
Applied Missing Data Analysis in the Health Sciences



XIAO-HUA ZHOU
CHUAN ZHOU
DANPING LIU
XIAOBO DING

STATISTICS IN PRACTICE

WILEY

APPLIED MISSING DATA ANALYSIS IN HEALTH SCIENCES

XIAO-HUA ZHOU
University of Washington

CHUAN ZHOU
University of Washington

DANPING LIU
National Institutes of Health

XIAOBO DING
Chinese Academy of Sciences

常州大学图书馆
藏书章

WILEY

Copyright 2014 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate percopy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, JohnWiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herin may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department with the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Zhou, Xiao-Hua, author.
Applied missing data analysis in the health sciences / Xiao-Hua Zhou, Chuan Zhou, Danping Liu, Xiaobo Ding.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-52381-0 (cloth)
I. Zhou, Chuan, 1972- author. II. Liu, Danping, 1981- author. III. Ding, Xiaobo, author. IV. Title.
[DNLM: 1. Data Interpretation, Statistical. 2. Biomedical Research–methods. 3. Models, Statistical. 4. Research Design. WA 950]
R853.S7
610.72′4–dc23

2013047830

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

APPLIED MISSING DATA ANALYSIS IN HEALTH SCIENCES

WILEY SERIES IN STATISTICS IN PRACTICE

Advisory Editor, Marian Scott, *University of Glasgow, Scotland, UK*

Founding Editor, Vic Barnett, *Nottingham Trent University, UK*

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 sound motivation 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 employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance, and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

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

试读结束，需要全本PDF请购买 www.ertongbook.com

To Yea-Jae, Yi,
Tingting, and Shuqin

PREFACE

With a strong practical emphasis on health science applications, this book describes statistical methods and models for the analysis of data with missing values. We attempt to write so that researchers with experience in applied data analysis, but less technical knowledge than a statistician, should be able to understand and implement most of the methods described. For those with a stronger background in statistics, we provide more technical details as to not detract from the flow of rest of the chapter. We have also tried to choose examples that are relevant to most health science researchers who work in a variety of disciplines.

In all fields of study, missing data are a common problem since, for any data collection process, there are so many things that could go wrong that missing values are all too likely. Thus, when attempts are made to answer the scientific questions of interest, researchers ask the all-too-common question: what do we do with the missing data?

The statistical literature to answer this question is well developed, but overly technical and complicated for researchers who are not experts in statistics and methodology. Therefore, researchers may recognize the existence of missing data, but fail to respond for two reasons: first, they may not understand the consequences of ignoring missing data and how it can impact the validity of their results; second, there is a lack of understanding of the statistical methods for missing data and how to apply them in their own research. Therefore, the purpose of this book is to provide health science researchers with the means of understanding the importance of missing data in their own personal research and the ability to use these methods given the available software.

This book is organized into eight chapters. Chapter 1 introduces concepts on the missing data mechanism and some real-world examples. Chapter 2 gives an overview of methods for dealing with missing data. Chapter 3 describes some design strategies for minimizing the impact of missing data. Chapters 4 and 5 introduce methods for dealing with missing data problems in cross-sectional and longitudinal studies, respectively. Chapter 6 deals exclusively with missing data problems in survival analysis. Whereas Chapters 3–6 deal with ignorable missing data problems, Chapter 7 presents methods for dealing with nonignorable missing data problems. Finally, Chapter 8 discusses methods for dealing with missing data in causal inferences.

As we worked through examples in the book, we chose to provide software code in the text of the chapters as we want to encourage application of these methods after an understanding of the basic theory. We chose to include R code in the text as many of the methods can be implemented in R; in addition, R is also a publicly available software environment (see www.r-project.org). Since many researchers also use Stata in addition to R, we include code for some selected examples. All the analysis data sets, together with R and Stata codes used in this book, can be downloaded from http://faculty.washington.edu/azhou/.

X.H. ZHOU, C. ZHOU, D. LIU, AND X. DING

Seattle, Washington
March, 2014

CONTENTS IN BRIEF

1 MISSING DATA CONCEPTS AND MOTIVATING EXAMPLES 1

2 OVERVIEW OF METHODS FOR DEALING WITH MISSING DATA 15

3 DESIGN CONSIDERATIONS IN THE PRESENCE OF MISSING DATA 25

4 CROSS-SECTIONAL DATA METHODS 31

5 LONGITUDINAL DATA METHODS 69

6 SURVIVAL ANALYSIS UNDER IGNORABLE MISSINGNESS 121

7 NONIGNORABLE MISSINGNESS 147

8 ANALYSIS OF RANDOMIZED CLINICAL TRIALS WITH NONCOMPLIANCE 185

BIBLIOGRAPHY 215

CONTENTS

LIST OF FIGURES xv

LIST OF TABLES xvii

PREFACE xix

1 MISSING DATA CONCEPTS AND MOTIVATING EXAMPLES 1

1.1 Overview of the Missing Data Problem / 1

1.2 Patterns and Mechanisms of Missing Data / 2

1.2.1 Missing Data Patterns / 3

1.2.2 Missing Data Mechanisms / 5

1.3 Data Examples / 7

1.3.1 Improving Mood and Promoting Access to Collaborative Treatment (IMPACT) Study / 9

1.3.2 National Alzheimer's Coordinating Center Minimum Data Set / 10

1.3.3 National Alzheimer's Coordinating Center Uniform Data Set / 11

1.3.4 The Pathways Study / 12

1.3.5 Randomized Trial on Vitamin A Supplement / 12

1.3.6 Randomized Trial on Effectiveness of Flu Shot / 12

2 OVERVIEW OF METHODS FOR DEALING WITH MISSING DATA 15

2.1 Methods That Remove Observations / 15

2.1.1 Complete-Case Methods / 16
2.1.2 Weighted Complete-Case Methods / 16
2.1.3 Removing Variables with Large Amounts of Missing Values / 17
2.2 Methods That Utilize All Available Data / 17
2.2.1 Maximum Likelihood / 18
2.3 Methods That Impute Missing Values / 19
2.3.1 Single Imputation Methods / 19
2.3.2 Multiple Imputation / 21
2.4 Bayesian Methods / 24

3 DESIGN CONSIDERATIONS IN THE PRESENCE OF MISSING DATA 25
3.1 Design Factors Related to Missing Data / 25
3.2 Strategies for Limiting Missing Data in the Design of Clinical Trials / 27
3.3 Strategies for Limiting Missing Data in the Conduct of Clinical Trials / 28
3.4 Minimize the Impact of Missing Data / 29

4 CROSS-SECTIONAL DATA METHODS 31
4.1 Overview of General Methods / 31
4.2 Data Examples / 31
4.2.1 Simulation Study / 32
4.2.2 NHANES Example / 32
4.3 Maximum Likelihood Approach / 32
4.3.1 EM Algorithm for Linear Regression with a Missing Continuous Covariate / 34
4.3.2 EM Algorithm for Linear Regression with Missing Discrete Covariate / 36
4.3.3 EM Algorithm for Logistic Regression with Missing Binary Outcome / 38
4.3.4 Simulation Study / 38
4.3.5 IMPACT Study / 39
4.3.6 NACC Study / 40
4.4 Bayesian Methods / 40
4.4.1 Theory / 40
4.4.2 Joint Model and Ignorable Missingness / 41
4.4.3 Bayesian Computation for Missing Data / 43
4.4.4 Simulation Example / 44

试读结束，需要全本PDF请购买 www.ertongbook.com

4.4.5 IMPACT Study / 45
4.4.6 NHANES Example / 45
4.5 Multiple Imputation / 47
4.5.1 Theory / 47
4.5.2 Some General Guidelines on Imputation Models and Analysis Models / 49
4.5.3 Theoretical Justification for the MI Method / 50
4.5.4 MI When θ Is k-Dimensional / 51
4.5.5 Simulated Example / 53
4.5.6 IMPACT Study / 53
4.6 Imputing Estimating Equations / 54
4.7 Inverse Probability Weighting / 55
4.7.1 Theory / 55
4.7.2 Simulated Example / 56
4.8 Doubly Robust Estimators / 56
4.8.1 Theory / 56
4.8.2 Variance Estimation / 57
4.8.3 NACC Study / 59
4.9 Code Used in This Chapter / 60
4.9.1 Code Used in Section 4.3.4 / 60
4.9.2 Code Used in Section 4.3.5 / 62
4.9.3 Code Used in Section 4.4.4 / 63
4.9.4 Code Used in Section 4.4.5 / 64
4.9.5 Code Used in Section 4.4.6 / 65
4.9.6 Code Used in Section 4.5.5 / 66
4.9.7 Code Used in Section 4.5.6 / 66
4.9.8 Code Used in Section 4.7.2 / 67
5 LONGITUDINAL DATA METHODS **69**
5.1 Overview / 69
5.2 Examples / 70
5.2.1 IMPACT Study / 70
5.2.2 NACC UDS Data / 71
5.3 Longitudinal Regression Models for Complete Data / 73
5.3.1 Linear Mixed Models for Continuous Longitudinal Data / 73
5.3.2 Generalized Estimating Equations / 77
5.3.3 Generalized Linear Mixed Models / 79
5.3.4 Time-Dependent Covariates / 81
5.4 Missing Data Settings and Simple Methods / 82

5.4.1 Setup / 82
5.4.2 Simple Methods / 82
5.5 Likelihood Approach / 83
5.5.1 Example: IMPACT Study / 84
5.6 Inverse Probability Weighted GEE with MAR Dropout / 86
5.6.1 Modeling the Selection Probability / 86
5.6.2 IPWGEE1 and IPWGEE2 / 87
5.6.3 A Simulation Study / 89
5.6.4 Example: IMPACT Study / 90
5.7 Extension to Nonmonotone Missingness / 91
5.8 Multiple Imputation / 91
5.8.1 Joint Imputation Model / 92
5.8.2 Imputation by Chained Equations / 93
5.8.3 Example: NACC UDS Data / 94
5.8.4 Example: IMPACT Study / 96
5.9 Bayesian Inference / 99
5.10 Other Approaches / 101
5.10.1 Imputing Estimating Equations / 101
5.10.2 Doubly Robust Estimation / 103
5.10.3 Missing Outcome and Covariates / 104
Appendix 5.A: Technical Details of the Approximation Methods for GLMM and Computer Code for the Examples / 109
5.A.1 PQL and MQL / 109
5.A.2 Laplace Approximation / 110
5.A.3 Gaussian Quadrature / 111
5.A.4 Code for This Chapter / 111

6 SURVIVAL ANALYSIS UNDER IGNORABLE MISSINGNESS 121

6.1 Overview / 121
6.2 Introduction / 122
6.2.1 Review of the Cox Model with Completely Observed Covariates / 122
6.2.2 Missing Data / 124
6.3 Enhanced Complete-Case Analysis / 125
6.4 Weighted Methods / 127
6.4.1 Simple Weighted Estimations / 127
6.4.2 Augmented Weighted Estimation / 131
6.4.3 Reweighting Estimators / 134
6.5 Imputation Methods / 136

6.5.1 Imputation by Conditional Expectations / 136

6.5.2 Multiple Imputation / 137

6.6 Nonparametric Maximum Likelihood Estimation / 139

6.7 Transformation Model / 140

6.8 Data Example: Pathways Study / 142

6.9 Concluding Remarks / 144

7 NONIGNORABLE MISSINGNESS 147

7.1 Introduction / 147

7.2 Cross-Sectional Data: Selection Model / 149

7.2.1 Missing Outcome / 149

7.2.2 Missing Covariate / 150

7.3 Longitudinal Data with Dropout / 150

7.3.1 Notation / 150

7.3.2 Overview / 151

7.3.3 Pattern Mixture Model / 151

7.3.4 Selection Model / 155

7.3.5 Shared Random Effects Model / 157

7.3.6 Mixed Effects Hybrid Model / 159

7.3.7 Comparison of Likelihood-Based Methods / 160

7.4 Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates / 161

7.4.1 Generalized Linear Models / 162

7.4.2 Modeling the Covariate Distribution / 163

7.4.3 Modeling the Missing Data Mechanism / 163

7.4.4 Prior Distributions / 164

7.4.5 Posterior Computation / 165

7.5 Multiple Imputation / 165

7.5.1 Proper Multiple Imputation by Bayesian Analysis / 166

7.5.2 Approximate Bayesian Bootstrap Hot Deck Imputation / 167

7.6 Inverse Probability Weighted Methods / 170

7.6.1 Semiparametric Regression for Repeated Outcomes / 170

7.6.2 Estimation of Marginal Mean of an Outcome from Longitudinal Data / 174

8 ANALYSIS OF RANDOMIZED CLINICAL TRIALS WITH NONCOMPLIANCE 185

8.1 Overview / 185

8.2 Examples / 186

8.2.1 Randomized Trial on Vitamin A Supplement / 186

8.2.2 Randomized Trial on Effectiveness of Flu Shot / 187

8.3 Some Common but Naive Methods / 188

8.4 Notations, Assumptions, and Causal Definitions / 189

8.5 Method of Instrumental Variables / 192

8.6 Moment-Based Method / 194

8.7 Maximum Likelihood and Bayesian Methods / 197

8.8 Noncompliance and Missing Outcome Data / 202

8.9 Analysis of the Two Examples / 211

8.9.1 Analysis of Vitamin A Supplement Example / 211

8.9.2 Flu Shot Example / 212

8.10 Other Methods for Dealing with Both Noncompliance and Missing Data / 212

Appendix 8.A: Multivariate Delta Method / 213

BIBLIOGRAPHY **215**

INDEX **225**

LIST OF FIGURES

1.1 Example 1.6 to explain the missing mechanisms, MCAR, MAR, and MNAR 8

4.1 Scatter plot of height versus weight 39

5.1 Forty individuals and mean trajectories 70

5.2 Diagnostic plot for the imputed `hachin`, `education`, `mmse2`, and `apoee4` 96

5.3 Illustration of the imputation procedures 103

LIST OF TABLES

1.1 Regression Results Under Different Missing Data Mechanisms, Using Full Data, Data with Completely Observed Values, and Data with Missing Values, Respectively 7

1.2 Missing Data Pattern for SCL-20 Scores Over Time 10

4.1 Linear Regression with Y_2 as Outcome 39

4.2 Linear Regression with Y_2 as Covariate 39

4.3 Cross-Sectional Analysis Result for the IMPACT study 40

4.4 Cross-Sectional Analysis Result for the NACC Study 41

4.5 Bayesian Analysis Results for the NHANES Example 47

4.6 Regression Analysis for the Cross-Sectional Simulated Data Based on Imputation Methods 53

4.7 The Estimated Location and Scale Parameters with the Associated Standard Errors for the NACC Data 59

5.1 The Missing Data Pattern in the SCL-20 Scores at Baseline, 3, 6, 12, 18, and 24 Months 72

5.2 Missing Data Pattern for the UDS Data 73

5.3 Common Choices of Working Correlation R_i with the Cluster Size $n_i = 4$ 78

试读结束，需要全本PDF请购买 www.ertongbook.com