

STATISTICAL ANALYSIS WITH MISSING DATA

SECOND EDITION

**Roderick J. A. Little
& Donald B. Rubin**

WILEY SERIES IN
PROBABILITY AND STATISTICS

Statistical Analysis with Missing Data

Second Edition

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie,
Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louis M. Ryan,
David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*

Editors Emeriti: *Vic Barnett, J. Stuart Hunder, David G. Kendall*

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

Statistical Analysis with Missing Data

Second Edition

RODERICK J. A. LITTLE

DONALD B. RUBIN



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2002 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 Sections 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 per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, e-mail: permcoordinator@wiley.com.

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 herein 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 within 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

Little, Roderick J. A.

Statistical analysis with missing data / Roderick J Little, Donald B. Rubin. -- 2nd ed.

p. cm. -- (Wiley series in probability and statistics)

"A Wiley-Interscience publication."

Includes bibliographical references and index.

ISBN 0-471-18386-5 (acid-free paper)

1. Mathematical statistics. 2. Missing observations (Statistics) I Rubin, Donald B. II.

Title. III. Series

QA276 .L57 2002

519.5--dc21

2002027006

ISBN 0-471-18386-5

Printed in the United States of America.

20 19 18 17 16 15 14

Contents

Preface	xiii
----------------	-------------

PART I OVERVIEW AND BASIC APPROACHES

1. Introduction	3
1.1. The Problem of Missing Data, 3	
1.2. Missing-Data Patterns, 4	
1.3. Mechanisms That Lead to Missing Data, 11	
1.4. A Taxonomy of Missing-Data Methods, 19	
2. Missing Data in Experiments	24
2.1. Introduction, 24	
2.2. The Exact Least Squares Solution with Complete Data, 25	
2.3. The Correct Least Squares Analysis with Missing Data, 27	
2.4. Filling in Least Squares Estimates, 28	
2.4.1. Yates's Method, 28	
2.4.2. Using a Formula for the Missing Values, 29	
2.4.3. Iterating to Find the Missing Values, 29	
2.4.4. ANCOVA with Missing-Value Covariates, 30	
2.5. Bartlett's ANCOVA Method, 30	
2.5.1. Useful Properties of Bartlett's Method, 30	
2.5.2. Notation, 30	
2.5.3. The ANCOVA Estimates of Parameters and Missing Y Values, 31	
2.5.4. ANCOVA Estimates of the Residual Sums of Squares and the Covariance Matrix of $\hat{\beta}$, 31	

2.6.	Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods, 33	
2.7.	Correct Least Squares Estimates of Standard Errors and One Degree of Freedom Sums of Squares, 35	
2.8.	Correct Least Squares Sums of Squares with More Than One Degree of Freedom, 37	
3.	Complete-Case and Available-Case Analysis, Including Weighting Methods	41
3.1.	Introduction, 41	
3.2.	Complete-Case Analysis, 41	
3.3.	Weighted Complete-Case Analysis, 44	
3.3.1.	Weighting Adjustments, 44	
3.3.2.	Added Variance from Nonresponse Weighting, 50	
3.3.3.	Post-Stratification and Raking To Known Margins, 51	
3.3.4.	Inference from Weighted Data, 53	
3.3.5.	Summary of Weighting Methods, 53	
3.4.	Available-Case Analysis, 53	
4.	Single Imputation Methods	59
4.1.	Introduction, 59	
4.2.	Imputing Means from a Predictive Distribution, 61	
4.2.1.	Unconditional Mean Imputation, 61	
4.2.2.	Conditional Mean Imputation, 62	
4.3.	Imputing Draws from a Predictive Distribution, 64	
4.3.1.	Draws Based on Explicit Models, 64	
4.3.2.	Draws Based on Implicit Models, 66	
4.4.	Conclusions, 72	
5.	Estimation of Imputation Uncertainty	75
5.1.	Introduction, 75	
5.2.	Imputation Methods that Provide Valid Standard Errors from a Single Filled-in Data Set, 76	
5.3.	Standard Errors for Imputed Data by Resampling, 79	
5.3.1.	Bootstrap Standard Errors, 79	
5.3.2.	Jackknife Standard Errors, 81	
5.4.	Introduction to Multiple Imputation, 85	
5.5.	Comparison of Resampling Methods and Multiple Imputation, 89	

PART II LIKELIHOOD-BASED APPROACHES TO THE ANALYSIS OF MISSING DATA

6. Theory of Inference Based on the Likelihood Function	97
6.1. Review of Likelihood-Based Estimation for Complete Data, 97	
6.1.1. Maximum Likelihood Estimation, 97	
6.1.2. Rudiments of Bayes Estimation, 104	
6.1.3. Large-Sample Maximum Likelihood and Bayes Inference, 105	
6.1.4. Bayes Inference Based on the Full Posterior Distribution, 112	
6.1.5. Simulating Draws from Posterior Distributions, 115	
6.2. Likelihood-Based Inference with Incomplete Data, 117	
6.3. A Generally Flawed Alternative to Maximum Likelihood: Maximizing Over the Parameters and the Missing Data, 124	
6.3.1. The Method, 124	
6.3.2. Background, 124	
6.3.3. Examples, 125	
6.4. Likelihood Theory for Coarsened Data, 127	
7. Factored Likelihood Methods, Ignoring the Missing-Data Mechanism	133
7.1. Introduction, 133	
7.2. Bivariate Normal Data with One Variable Subject to Nonresponse: ML Estimation, 133	
7.2.1. ML Estimates, 135	
7.2.2. Large-Sample Covariance Matrix, 139	
7.3. Bivariate Normal Monotone Data: Small-Sample Inference, 140	
7.4. Monotone Data With More Than Two Variables, 143	
7.4.1. Multivariate Data With One Normal Variable Subject to Nonresponse, 143	
7.4.2. Factorization of the Likelihood for a General Monotone Pattern, 144	
7.4.3. Computation for Monotone Normal Data via the Sweep Operator, 148	
7.4.4. Bayes Computation for Monotone Normal Data via the Sweep Operator, 155	
7.5. Factorizations for Special Nonmonotone Patterns, 156	

8. Maximum Likelihood for General Patterns of Missing Data: Introduction and Theory with Ignorable Nonresponse	164
8.1. Alternative Computational Strategies, 164	
8.2. Introduction to the EM Algorithm, 166	
8.3. The E and M Steps of EM, 167	
8.4. Theory of the EM Algorithm, 172	
8.4.1. Convergence Properties, 172	
8.4.2. EM for Exponential Families, 175	
8.4.3. Rate of Convergence of EM, 177	
8.5. Extensions of EM, 179	
8.5.1. ECM Algorithm, 179	
8.5.2. ECME and AECM Algorithms, 183	
8.5.3. PX-EM Algorithm, 184	
8.6. Hybrid Maximization Methods, 186	
 9. Large-Sample Inference Based on Maximum Likelihood Estimates	 190
9.1. Standard Errors Based on the Information Matrix, 190	
9.2. Standard Errors via Methods that do not Require Computing and Inverting an Estimate of the Observed Information Matrix, 191	
9.2.1. Supplemental EM Algorithm, 191	
9.2.2. Bootstrapping the Observed Data, 196	
9.2.3. Other Large Sample Methods, 197	
9.2.4. Posterior Standard Errors from Bayesian Methods, 198	
 10. Bayes and Multiple Imputation	 200
10.1. Bayesian Iterative Simulation Methods, 200	
10.1.1. Data Augmentation, 200	
10.1.2. The Gibbs' Sampler, 203	
10.1.3. Assessing Convergence of Iterative Simulations, 206	
10.1.4. Some Other Simulation Methods, 208	
10.2. Multiple Imputation, 209	
10.2.1. Large-Sample Bayesian Approximation of the Posterior Mean and Variance Based on a Small Number of Draws, 209	
10.2.2. Approximations Using Test Statistics, 212	
10.2.3. Other Methods for Creating Multiple Imputations, 214	

PART III LIKELIHOOD-BASED APPROACHES TO THE ANALYSIS OF INCOMPLETE DATA: SOME EXAMPLES

11. Multivariate Normal Examples, Ignoring the Missing-Data Mechanism	223
11.1. Introduction, 223	
11.2. Inference for a Mean Vector and Covariance Matrix with Missing Data Under Normality, 223	
11.2.1. The EM Algorithm for Incomplete Multivariate Normal Samples, 226	
11.2.2. Estimated Asymptotic Covariance Matrix of $(\theta - \hat{\theta})$, 226	
11.2.3. Bayes Inference for the Normal Model via Data Augmentation, 227	
11.3. Estimation with a Restricted Covariance Matrix, 231	
11.4. Multiple Linear Regression, 237	
11.4.1. Linear Regression with Missing Values Confined to the Dependent Variable, 237	
11.4.2. More General Linear Regression Problems with Missing Data, 239	
11.5. A General Repeated-Measures Model with Missing Data, 241	
11.6. Time Series Models, 246	
11.6.1. Introduction, 246	
11.6.2. Autoregressive Models for Univariate Time Series with Missing Values, 246	
11.6.3. Kalman Filter Models, 248	
12. Robust Estimation	253
12.1. Introduction, 253	
12.2. Robust Estimation for a Univariate Sample, 253	
12.3. Robust Estimation of the Mean and Covariance Matrix, 255	
12.3.1. Multivariate Complete Data, 255	
12.3.2. Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values, 257	
12.3.3. Adaptive Robust Multivariate Estimation, 259	
12.3.4. Bayes Inferences for the t Model, 259	
12.4. Further Extensions of the t Model, 260	
13. Models for Partially Classified Contingency Tables, Ignoring the Missing-Data Mechanism	266
13.1. Introduction, 266	

13.2.	Factored Likelihoods for Monotone Multinomial Data, 267	
13.2.1.	Introduction, 267	
13.2.2.	ML Estimation for Monotone Patterns, 268	
13.2.3.	Precision of Estimation, 275	
13.3.	ML and Bayes Estimation for Multinomial Samples with General Patterns of Missing Data, 278	
13.4.	Loglinear Models for Partially Classified Contingency Tables, 281	
13.4.1.	The Complete-Data Case, 281	
13.4.2.	Loglinear Models for Partially Classified Tables, 285	
13.4.3.	Goodness-of-Fit Tests for Partially Classified Data, 289	
14.	Mixed Normal and Non-normal Data with Missing Values, Ignoring the Missing-Data Mechanism	292
14.1.	Introduction, 292	
14.2.	The General Location Model, 292	
14.2.1.	The Complete-Data Model and Parameter Estimates, 292	
14.2.2.	ML Estimation with Missing Values, 294	
14.2.3.	Details of the E Step Calculations, 296	
14.2.4.	Bayes Computations for the Unrestricted General Location Model, 298	
14.3.	The General Location Model with Parameter Constraints, 300	
14.3.1.	Introduction, 300	
14.3.2.	Restricted Models for the Cell Means, 300	
14.3.3.	Loglinear Models for the Cell Probabilities, 303	
14.3.4.	Modifications to the Algorithms of Sections 14.2.2 and 14.2.3 for Parameter Restrictions, 303	
14.3.5.	Simplifications when the Categorical Variables are More Observed than the Continuous Variables, 305	
14.4.	Regression Problems Involving Mixtures of Continuous and Categorical Variables, 306	
14.4.1.	Normal Linear Regression with Missing Continuous or Categorical Covariates, 306	
14.4.2.	Logistic Regression with Missing Continuous or Categorical Covariates, 308	
14.5.	Further Extensions of the General Location Model, 309	
15.	Nonignorable Missing-Data Models	312
15.1.	Introduction, 312	

15.2.	Likelihood Theory for Nonignorable Models,	315
15.3.	Models with Known Nonignorable Missing-Data Mechanisms: Grouped and Rounded Data,	316
15.4.	Normal Selection Models,	321
15.5.	Normal Pattern-Mixture Models,	327
15.5.1.	Univariate Normal Pattern-Mixture Models,	327
15.5.2.	Bivariate Normal Pattern-Mixture Models Identified via Parameter Restrictions,	331
15.6.	Nonignorable Models for Normal Repeated-Measures Data,	336
15.7.	Nonignorable Models for Categorical Data,	340
References		349
Author Index		365
Subject Index		371

Preface

The literature on the statistical analysis of data with missing values has flourished since the early 1970s, spurred by advances in computer technology that made previously laborious numerical calculations a simple matter. This book aims to survey current methodology for handling missing-data problems and present a likelihood-based theory for analysis with missing data that systematizes these methods and provides a basis for future advances. Part I of the book discusses historical approaches to missing-value problems in three important areas of statistics: analysis of variance of planned experiments, survey sampling, and multivariate analysis. These methods, although not without value, tend to have an ad hoc character, often being solutions worked out by practitioners with limited research into theoretical properties. Part II presents a systematic approach to the analysis of data with missing values, where inferences are based on likelihoods derived from formal statistical models for the data-generating and missing-data mechanisms. Part III presents applications of the approach in a variety of contexts, including ones involving regression, factor analysis, contingency table analysis, time series, and sample survey inference. Many of the historical methods in Part I can be derived as examples (or approximations) of this likelihood-based approach.

The book is intended for the applied statistician and hence emphasizes examples over the precise statement of regularity conditions or proofs of theorems. Nevertheless, readers are expected to be familiar with basic principles of inference based on likelihoods, briefly reviewed in Section 6.1. The book also assumes an understanding of standard models of complete-data analysis—the normal linear model, multinomial models for counted data—and the properties of standard statistical distributions, especially the multivariate normal distribution. Some chapters assume familiarity in particular areas of statistical activity—analysis of variance for experimental designs (Chapter 2), survey sampling (Chapters 3, 4, and 5), or loglinear models for contingency tables (Chapter 13). Specific examples also introduce other statistical topics, such as factor analysis or time series (Chapter 11). The discussion of these examples is self-contained and does not require specialized knowledge, but such knowledge will, of course, enhance the reader's appreciation