

Statistical Analysis with Missing Data

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by *Walter A. Shewhart and Samuel S. Wilks*

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, Jozef L. Teugels*

The **Wiley Series in Probability and Statistics** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at

<http://www.wiley.com/go/wsp>

Statistical Analysis with Missing Data

Roderick J. A. Little

Richard D. Remington Distinguished University Professor of Biostatistics, Professor of Statistics, and Research Professor, Institute for Social Research, at the University of Michigan

Donald B. Rubin

Professor at Yau Mathematical Sciences Center, Tsinghua University; Murray Shusterman Senior Research Fellow, Fox School of Business, at Temple University; and Professor Emeritus, at Harvard University

3rd Edition

WILEY

This edition first published 2020
© 2020 John Wiley & Sons, Inc

Edition History

All rights reserved. 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 or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Roderick J A Little and Donald B Rubin to be identified as the authors of the material in this work has been asserted in accordance with law.

Registered Office(s)

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

In view of ongoing research, equipment modifications, changes in governmental regulations, and the constant flow of information relating to the use of experimental reagents, equipment, and devices, the reader is urged to review and evaluate the information provided in the package insert or instructions for each chemical, piece of equipment, reagent, or device for, among other things, any changes in the instructions or indication of usage and for added warnings and precautions. While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Little, Roderick J.A., author. | Rubin, Donald B., author.

Title: Statistical analysis with missing data / Roderick J.A. Little, Donald B. Rubin.

Description: Third edition | Hoboken, NJ : Wiley, 2020. | Series: Wiley series in probability and statistics | Includes index. |

Identifiers: LCCN 2018058860 (print) | LCCN 2018061330 (ebook) | ISBN 9781118596012 (Adobe PDF) | ISBN 9781118595695 (ePub) | ISBN 9780470526798 (hardcover)

Subjects: LCSH: Mathematical statistics. | Mathematical statistics—Problems, exercises, etc. | Missing observations (Statistics) | Missing observations (Statistics)—Problems, exercises, etc.

Classification: LCC QA276 (ebook) | LCC QA276 .L57 2019 (print) | DDC 519.5—dc23

LC record available at <https://lccn.loc.gov/2018058860>

Cover image: Wiley

Cover design by Wiley

Set in 10/12pt WarnockPro by Aptara Inc., New Delhi, India

Contents

Preface to the Third Edition *xi*

Part I Overview and Basic Approaches 1

1	Introduction 3
1.1	The Problem of Missing Data 3
1.2	Missingness Patterns and Mechanisms 8
1.3	Mechanisms That Lead to Missing Data 13
1.4	A Taxonomy of Missing Data Methods 23
2	Missing Data in Experiments 29
2.1	Introduction 29
2.2	The Exact Least Squares Solution with Complete Data 30
2.3	The Correct Least Squares Analysis with Missing Data 32
2.4	Filling in Least Squares Estimates 33
2.4.1	Yates's Method 33
2.4.2	Using a Formula for the Missing Values 34
2.4.3	Iterating to Find the Missing Values 34
2.4.4	ANCOVA with Missing Value Covariates 35
2.5	Bartlett's ANCOVA Method 35
2.5.1	Useful Properties of Bartlett's Method 35
2.5.2	Notation 36
2.5.3	The ANCOVA Estimates of Parameters and Missing Y -Values 36
2.5.4	ANCOVA Estimates of the Residual Sums of Squares and the Covariance Matrix of $\hat{\beta}$ 37
2.6	Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods 38
2.7	Correct Least Squares Estimates of Standard Errors and One Degree of Freedom Sums of Squares 40

2.8	Correct Least-Squares Sums of Squares with More Than One Degree of Freedom	42
3	Complete-Case and Available-Case Analysis, Including Weighting Methods	47
3.1	Introduction	47
3.2	Complete-Case Analysis	47
3.3	Weighted Complete-Case Analysis	50
3.3.1	Weighting Adjustments	50
3.3.2	Poststratification and Raking to Known Margins	58
3.3.3	Inference from Weighted Data	60
3.3.4	Summary of Weighting Methods	61
3.4	Available-Case Analysis	61
4	Single Imputation Methods	67
4.1	Introduction	67
4.2	Imputing Means from a Predictive Distribution	69
4.2.1	Unconditional Mean Imputation	69
4.2.2	Conditional Mean Imputation	70
4.3	Imputing Draws from a Predictive Distribution	73
4.3.1	Draws Based on Explicit Models	73
4.3.2	Draws Based on Implicit Models – Hot Deck Methods	76
4.4	Conclusion	81
5	Accounting for Uncertainty from Missing Data	85
5.1	Introduction	85
5.2	Imputation Methods that Provide Valid Standard Errors from a Single Filled-in Data Set	86
5.3	Standard Errors for Imputed Data by Resampling	90
5.3.1	Bootstrap Standard Errors	90
5.3.2	Jackknife Standard Errors	92
5.4	Introduction to Multiple Imputation	95
5.5	Comparison of Resampling Methods and Multiple Imputation	100

Part II Likelihood-Based Approaches to the Analysis of Data with Missing Values 107

6	Theory of Inference Based on the Likelihood Function	109
6.1	Review of Likelihood-Based Estimation for Complete Data	109
6.1.1	Maximum Likelihood Estimation	109
6.1.2	Inference Based on the Likelihood	118
6.1.3	Large Sample Maximum Likelihood and Bayes Inference	119

6.1.4	Bayes Inference Based on the Full Posterior Distribution	126
6.1.5	Simulating Posterior Distributions	130
6.2	Likelihood-Based Inference with Incomplete Data	132
6.3	A Generally Flawed Alternative to Maximum Likelihood: Maximizing over the Parameters and the Missing Data	141
6.3.1	The Method	141
6.3.2	Background	142
6.3.3	Examples	143
6.4	Likelihood Theory for Coarsened Data	145
7	Factored Likelihood Methods When the Missingness Mechanism Is Ignorable	151
7.1	Introduction	151
7.2	Bivariate Normal Data with One Variable Subject to Missingness: ML Estimation	153
7.2.1	ML Estimates	153
7.2.2	Large-Sample Covariance Matrix	157
7.3	Bivariate Normal Monotone Data: Small-Sample Inference	158
7.4	Monotone Missingness with More Than Two Variables	161
7.4.1	Multivariate Data with One Normal Variable Subject to Missingness	161
7.4.2	The Factored Likelihood for a General Monotone Pattern	162
7.4.3	ML Computation for Monotone Normal Data via the Sweep Operator	166
7.4.4	Bayes Computation for Monotone Normal Data via the Sweep Operator	174
7.5	Factored Likelihoods for Special Nonmonotone Patterns	175
8	Maximum Likelihood for General Patterns of Missing Data: Introduction and Theory with Ignorable Nonresponse	185
8.1	Alternative Computational Strategies	185
8.2	Introduction to the EM Algorithm	187
8.3	The E Step and The M Step of EM	188
8.4	Theory of the EM Algorithm	193
8.4.1	Convergence Properties of EM	193
8.4.2	EM for Exponential Families	196
8.4.3	Rate of Convergence of EM	198
8.5	Extensions of EM	200
8.5.1	The ECM Algorithm	200
8.5.2	The ECME and AECM Algorithms	205
8.5.3	The PX-EM Algorithm	206
8.6	Hybrid Maximization Methods	208

9	Large-Sample Inference Based on Maximum Likelihood Estimates	213
9.1	Standard Errors Based on The Information Matrix	213
9.2	Standard Errors via Other Methods	214
9.2.1	The Supplemented EM Algorithm	214
9.2.2	Bootstrapping the Observed Data	219
9.2.3	Other Large-Sample Methods	220
9.2.4	Posterior Standard Errors from Bayesian Methods	221
10	Bayes and Multiple Imputation	223
10.1	Bayesian Iterative Simulation Methods	223
10.1.1	Data Augmentation	223
10.1.2	The Gibbs' Sampler	226
10.1.3	Assessing Convergence of Iterative Simulations	230
10.1.4	Some Other Simulation Methods	231
10.2	Multiple Imputation	232
10.2.1	Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws	232
10.2.2	Approximations Using Test Statistics or p -Values	235
10.2.3	Other Methods for Creating Multiple Imputations	238
10.2.4	Chained-Equation Multiple Imputation	241
10.2.5	Using Different Models for Imputation and Analysis	243
Part III Likelihood-Based Approaches to the Analysis of Incomplete Data: Some Examples 247		
11	Multivariate Normal Examples, Ignoring the Missingness Mechanism	249
11.1	Introduction	249
11.2	Inference for a Mean Vector and Covariance Matrix with Missing Data Under Normality	249
11.2.1	The EM Algorithm for Incomplete Multivariate Normal Samples	250
11.2.2	Estimated Asymptotic Covariance Matrix of $(\theta - \hat{\theta})$	252
11.2.3	Bayes Inference and Multiple Imputation for the Normal Model	253
11.3	The Normal Model with a Restricted Covariance Matrix	257
11.4	Multiple Linear Regression	264
11.4.1	Linear Regression with Missingness Confined to the Dependent Variable	264
11.4.2	More General Linear Regression Problems with Missing Data	266
11.5	A General Repeated-Measures Model with Missing Data	269

11.6	Time Series Models	273
11.6.1	Introduction	273
11.6.2	Autoregressive Models for Univariate Time Series with Missing Values	273
11.6.3	Kalman Filter Models	276
11.7	Measurement Error Formulated as Missing Data	277
12	Models for Robust Estimation	285
12.1	Introduction	285
12.2	Reducing the Influence of Outliers by Replacing the Normal Distribution by a Longer-Tailed Distribution	286
12.2.1	Estimation for a Univariate Sample	286
12.2.2	Robust Estimation of the Mean and Covariance Matrix with Complete Data	288
12.2.3	Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values	290
12.2.4	Adaptive Robust Multivariate Estimation	291
12.2.5	Bayes Inference for the t Model	292
12.2.6	Further Extensions of the t Model	294
12.3	Penalized Spline of Propensity Prediction	298
13	Models for Partially Classified Contingency Tables, Ignoring the Missingness Mechanism	301
13.1	Introduction	301
13.2	Factored Likelihoods for Monotone Multinomial Data	302
13.2.1	Introduction	302
13.2.2	ML and Bayes for Monotone Patterns	303
13.2.3	Precision of Estimation	312
13.3	ML and Bayes Estimation for Multinomial Samples with General Patterns of Missingness	313
13.4	Loglinear Models for Partially Classified Contingency Tables	317
13.4.1	The Complete-Data Case	317
13.4.2	Loglinear Models for Partially Classified Tables	320
13.4.3	Goodness-of-Fit Tests for Partially Classified Data	326
14	Mixed Normal and Nonnormal Data with Missing Values, Ignoring the Missingness Mechanism	329
14.1	Introduction	329
14.2	The General Location Model	329
14.2.1	The Complete-Data Model and Parameter Estimates	329
14.2.2	ML Estimation with Missing Values	331
14.2.3	Details of the E Step Calculations	334

14.2.4	Bayes' Computation for the Unrestricted General Location Model	335
14.3	The General Location Model with Parameter Constraints	337
14.3.1	Introduction	337
14.3.2	Restricted Models for the Cell Means	340
14.3.3	Loglinear Models for the Cell Probabilities	340
14.3.4	Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions	340
14.3.5	Simplifications When Categorical Variables are More Observed than Continuous Variables	343
14.4	Regression Problems Involving Mixtures of Continuous and Categorical Variables	344
14.4.1	Normal Linear Regression with Missing Continuous or Categorical Covariates	344
14.4.2	Logistic Regression with Missing Continuous or Categorical Covariates	346
14.5	Further Extensions of the General Location Model	347
15	Missing Not at Random Models	351
15.1	Introduction	351
15.2	Models with Known MNAR Missingness Mechanisms: Grouped and Rounded Data	355
15.3	Normal Models for MNAR Missing Data	362
15.3.1	Normal Selection and Pattern-Mixture Models for Univariate Missingness	362
15.3.2	Following up a Subsample of Nonrespondents	364
15.3.3	The Bayesian Approach	366
15.3.4	Imposing Restrictions on Model Parameters	369
15.3.5	Sensitivity Analysis	376
15.3.6	Subsample Ignorable Likelihood for Regression with Missing Data	379
15.4	Other Models and Methods for MNAR Missing Data	382
15.4.1	MNAR Models for Repeated-Measures Data	382
15.4.2	MNAR Models for Categorical Data	385
15.4.3	Sensitivity Analyses for Chained-Equation Multiple Imputations	391
15.4.4	Sensitivity Analyses in Pharmaceutical Applications	396
References		405
Author Index		429
Subject Index		437

Preface to the Third Edition

There has been tremendous growth in the literature on statistical methods for handling missing data, and associated software, since the publication of the second edition of “Statistical Analysis with Missing Data” in 2002. Attempting to cover this literature comprehensively would add excessively to the length of the book and also change its character. Therefore, our additions have focused mainly on work with which we have been associated and we can write about with some authority. The main changes from the second edition are as follows:

1. Concerning theory, we have changed the “obs” and “mis” notation for observed and missing data, which, though intuitive, caused some confusion because subscripting data by “obs” was not intended to imply conditioning on the pattern of observed values. We now use subscript (0) to denote observed values and subscript (1) to denote missing values, which is in fact similar to the notation employed by Rubin’s original (1976a) paper. We have also been more specific about assumptions for ignoring the missing data mechanism for likelihood-based/Bayesian analyses and asymptotic frequentist analysis; the latter involves changing missing data patterns in repeated analysis. These changes reflect material in Mealli and Rubin (2015). A definition of “partially missing at random” and ignorability for parameter subsets has been added, based on Little et al. (2016a).
2. Data previously termed “not missing at random” are now called “missing not at random,” which we think is clearer.
3. Applications place greater emphasis on multiple imputation rather than direct computation of the posterior distribution of parameters. This new emphasis reflects the expansion of flexible software for multiple imputation, which makes the method attractive to applied statisticians.
4. We have added a number of additional missing data applications to measurement error, disclosure limitation, robust inference, and clinical trial data.

5. Chapter 15, on missing not at random data, has been completely revamped, including a number of new applications to subsample regression and sensitivity analysis
6. A number of minor errors in the previous edition have been corrected, although (as in all books), some probably remain and other new ones may have crept in – for which we apologize.

The ideal of using a consistent notation across all chapters, avoiding the use of the same symbol to mean different concepts, proved too hard given the range of topics covered. However, we have tried to maintain a consistent notation within chapters, and defined new uses of common letters as they arise. We hope different uses of the same symbol across chapters is not too confusing, and welcome suggestions for improvements.

Part I

Overview and Basic Approaches

1

Introduction

1.1 The Problem of Missing Data

Standard statistical methods have been developed to analyze rectangular data sets. Traditionally, the rows of the data matrix represent units, also called cases, observations, or subjects depending on context, and the columns represent characteristics or variables measured for each unit. The entries in the data matrix are nearly always real numbers, either representing the values of essentially continuous variables, such as age and income, or representing categories of response, which may be ordered (e.g., level of education) or unordered (e.g., race, sex). This book concerns the analysis of such a data matrix when some of the entries in the matrix are not observed. For example respondents in a household survey may refuse to report income; in an industrial experiment, some results are missing because of mechanical failures unrelated to the experimental process; in an opinion survey, some individuals may be unable to express a preference for one candidate over another.

In the first two examples, it is natural to treat the values that are not observed as missing, in the sense that there are actual underlying values that would have been observed if survey techniques had been better or the industrial equipment had been better maintained. In the third example, however, it is less clear that a well-defined candidate preference has been masked by the nonresponse; thus, it is less natural to treat the unobserved values as missing. Instead, in this example, the lack of a response is essentially an additional point in the sample space of the variable being measured, which identifies a “no preference” or “don’t know” stratum of the population for that variable.

Older review articles on the statistical analysis of data with missing values include Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster et al. (1977), Little and Rubin (1983a), Little and Schenker (1994), and Little (1997). More recent literature includes books on the topic, such as Schafer (1997), van Buuren (2012), Carpenter and Kenward (2014), and Raghunathan (2015).

Part I considers basic approaches, including analysis of the complete cases and associated weighting methods, and methods that impute (that is fill in), the missing values. Part II considers more principled approaches based on statistical models and the associated likelihood function, and Part III provides applications of these methods. Our generally preferred philosophy of inference can be termed “calibrated Bayes,” where the inference is Bayesian, using models that yield inferences with good frequentist properties (Rubin 1984, 2019; Little 2006). For example, 95% Bayesian credibility intervals should have approximately 95% confidence coverage in repeated sampling from the population. The method of multiple imputation has such a Bayesian justification but can be used in conjunction with standard frequentist approaches to the complete-data inference.

Most statistical software packages allow the identification of nonrespondents by creating one or more special codes for those entries of the data matrix that are not observed. More than one code might be used to identify particular types of nonresponse, such as “don’t know,” or “refuse to answer,” or “out of legitimate range.” Some statistical software excludes units that have missing value codes for any of the variables involved in an analysis. This strategy, which is often termed a “complete-case analysis,” is generally inappropriate because the investigator is usually interested in making inferences about the entire target population, rather than about the portion of the target population that would provide responses on all relevant variables in the analysis. Our aim is to describe a collection of techniques that are more generally appropriate than complete-case analysis when missing entries in the data set mask the underlying values.

Definition 1.1 Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value.

When Definition 1.1 applies, it makes sense to consider analyses that effectively predict, or “impute” (that is, fill in), the unobserved values. If, on the other hand, Definition 1.1 does not apply, then imputing the unobserved values makes little sense, and an analysis that creates strata of the population defined by the pattern of observed data is more appropriate. Example 1.1 describes a situation with longitudinal data on obesity where Definition 1.1 clearly makes sense. Example 1.2 describes the case of a randomized experiment where it makes sense for one outcome variable (survival) but not for another (quality of life); and Example 1.3 describes a situation in opinion polling where Definition 1.1 may or may not make sense, depending on the specific setting.

Example 1.1 Nonresponse for a Binary Outcome Measured at Three Times Points. Woolson and Clarke (1984) analyze data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in schoolchildren. Table 1.1 summarizes the pattern of missing data in the data matrix. Five

Table 1.1 Example 1.1: data matrix for children in a survey summarized by the pattern of missing data: 1 = missing, 0 = observed

Pattern	Variables					No. of children with pattern
	Age	Sex	Weight 1	Weight 2	Weight 3	
A	0	0	0	0	0	1770
B	0	0	0	0	1	631
C	0	0	0	1	0	184
D	0	0	1	0	0	645
E	0	0	0	1	1	756
F	0	0	1	0	1	370
G	0	0	1	1	0	500

variables (sex, age, and obesity for three rounds of the survey) are recorded for 4856 units; sex and age are completely recorded, but the three obesity variables are sometimes missing, thereby generating six patterns of missingness. Because age is recorded in five categories and the obesity variables are binary, the data can be displayed as counts in a contingency table. Table 1.2 displays the data in this form, with missingness of obesity treated as a third category of the variable, where O = obese, N = not obese, and M = missing. Thus, the pattern MON denotes missing at the first round, obese at the second round, and not obese at the third round, and the other five patterns are defined analogously.

Woolson and Clarke analyze these data by fitting multinomial distributions over the $3^3 - 1 = 26$ response categories for each column in Table 1.2. That is missingness is regarded as defining strata of the population. We suspect that for these data, it makes good sense to regard the nonrespondents as having a true underlying value for the obesity variable. Hence, we would argue for treating the nonresponse categories as missing value indicators and estimating the joint distribution of the three dichotomous outcome variables from the partially missing data. Appropriate methods for handling such categorical data with missing values effectively impute the values of obesity that are not observed, as described in Chapter 12. The methods involve quite straightforward modifications of existing algorithms for categorical data analysis, which are now widely available in statistical software packages. For an analysis of these data that averages over patterns of missing data, see Ekholm and Skinner (1998).

Example 1.2 *Causal Effects of Treatments with Survival and Quality of Life Outcomes.* Consider a randomized experiment with two drug treatment conditions $T = 0$ or 1 for participants (units), and suppose that a primary outcome

Table 1.2 Example 1.1: number of children classified by population and relative weight category in three rounds of a survey

Response	Males					Females				
	Age group					Age group				
Category ^a	5–7	7–9	9–11	11–13	13–15	5–7	7–9	9–11	11–13	13–15
NNN	90	150	152	119	101	75	154	148	129	91
NNO	9	15	11	7	4	8	14	6	8	9
NON	3	8	8	8	2	2	13	10	7	5
NOO	7	8	10	3	7	4	19	8	9	3
ONN	0	8	7	13	8	2	2	12	6	6
ONO	1	9	7	4	0	2	6	0	2	0
OON	1	7	9	11	6	1	6	8	7	6
OOO	8	20	25	16	15	8	21	27	14	15
NNM	16	38	48	42	82	20	25	36	36	83
NOM	5	3	6	4	9	0	3	0	9	15
ONM	0	1	2	4	8	0	1	7	4	6
OOM	0	11	14	13	12	4	It	17	13	23
NMN	9	16	13	14	6	7	16	8	31	5
NMO	3	6	5	2	1	2	3	1	4	0
OMN	0	1	0	1	0	0	0	1	2	0
OMO	0	3	3	4	1	1	4	4	6	1
MNN	129	42	36	18	13	109	47	39	19	11
MNO	18	2	5	3	1	22	4	6	1	1
MON	6	3	4	3	2	7	1	7	2	2
MOO	13	13	3	1	2	24	8	13	2	3
NMM	32	45	59	82	95	23	47	53	58	89
OMM	5	7	17	24	23	5	7	16	37	32
MNM	33	33	31	23	34	27	23	25	21	43
MOM	11	4	9	6	12	5	5	9	1	15
MMN	70	55	40	37	15	65	39	23	23	14
MMO	24	14	9	14	3	19	13	8	10	5

^aNNN indicates not obese in 1977, 1979, and 1981; O indicates obese; and M indicates missing in a given year.

Source: Woolson and Clarke (1984). Reproduced with permission of John Wiley & Sons, Inc.

of the study is survived ($D = 0$) or dead ($D = 1$) at one year after randomization to treatment condition. Using a potential outcomes notation, let $D_i(1)$ denote the one-year survival status if participant i is assigned treatment 1, and $D_i(0)$ denote survival status if participant i is assigned treatment 0. The causal effect of treatment 1 relative to treatment 0 on survival for participant i is defined as $D_i(1) - D_i(0)$. Estimation of this causal effect can be considered a missing data problem, in that only one treatment can be assigned to each participant; therefore, $D_i(0)$ is unobserved (“missing”) for participants assigned treatment 1, and $D_i(1)$ is unobserved (“missing”) for participants assigned treatment 0. Individual causal effects are unobserved, but randomization allows for unbiased estimation of average causal effects for a sample or population (Rubin 1974), which can be estimated from this missing data perspective. The survival outcome under the treatment not received can be legitimately modeled as “missing data” in the sense of Definition 1.1 because one can consider what the survival outcome would have been under the treatment not assigned, even though this outcome is never observed. Other applications of this “potential outcomes” formulation for inference about causal effects include Rubin (1978a), Angrist et al. (1996), Barnard et al. (1998), Hirano et al. (2000), Frangakis and Rubin (1999, 2001, 2002), and Little et al. (2009).

Rubin (2000) discusses the more complex situation where a “quality-of-life health indicator” Y ($Y > 0$) is also measured as a secondary outcome for those still alive one year after randomization to treatment. For participants who die within a year of randomization, Y is undefined. The term “censored” due to death is sometimes used to describe this, but we think it usually makes little sense to treat these outcomes as missing values, in the sense defined by Definition 1.1, because quality of life is not a meaningful concept for people who are not alive. More specifically, let $D_i(T)$ denote the potential one-year survival outcome (1 = died, 0 = survived) for unit i under treatment T , as before. The potential outcomes on D can be used to classify the units into four strata as follows:

Those who would live under either treatment assignment, $LL = \{i | D_i(1) = D_i(0) = 0\}$

Those who would die under either treatment assignment, $DD = \{i | D_i(1) = D_i(0) = 1\}$

Those who would live under treatment but die under control, $LD = \{i | D_i(1) = 0, D_i(0) = 1\}$

Those who would die under treatment but live under control, $DL = \{i | D_i(1) = 1, D_i(0) = 0\}$.

For the LL units, there is a bivariate distribution of unit-level potential outcomes of Y under treatment and control, with one of these outcomes being observed and one missing. For the DD units, there is no information on Y , and

it is dubious to treat these values as missing. For the *LD* units, there is a distribution of Y under the treatment condition but not under the control condition, and for the *DL* units, there is a distribution of Y under the control condition but not under the treatment condition. Causal inference about Y can be conceptualized within this framework as imputing the survival status of units under the treatment not received and subsequently imputing quality of life of units under the treatment not received *within the stratum of LL units*.

Example 1.3 Nonresponse in Opinion Polls. Consider the situation where the units are individuals, who are polled about how they will vote in a future referendum, where the available responses are “yes,” “no,” or “missing.” Individuals who fail to respond to the question may be refusing to reveal real answers or may have no interest in voting. Definition 1.1 would not apply to individuals who would not vote, and these individuals define a stratum of the population that is not relevant to the outcome of the referendum, if only those who vote are considered. Definition 1.1 would apply to individuals who do not respond to the initial poll but would vote in the referendum; for these individuals, it makes sense to apply a method that effectively imputes a “yes” or “no” vote when analyzing the polling data. Rubin et al. (1996) consider a situation where there is a complete list of eligible voters, and persons who do not vote were counted as “nos” in the referendum. Here Definition 1.1 applies to all the unobserved values in the initial poll. Consequently, Rubin et al. (1996) consider methods that effectively impute the missing responses under a variety of modeling assumptions, as discussed in Example 15.19.

1.2 Missingness Patterns and Mechanisms

We find it useful to distinguish the missingness *pattern*, which describes which values are missing and observed in the data matrix and the missingness *mechanism* (or mechanisms), which concerns the relationship between missingness and the values of variables in the data matrix. Some methods of analysis, such as those described in Chapter 7, are intended for particular patterns of missing data and use only standard complete data analyses. Other methods, such as those described in Chapters 8–10, are applicable to more general missingness patterns, but usually involve more computing than methods designed for special patterns. Thus, it is beneficial to sort rows and columns of the data according to the missingness pattern to see if an orderly pattern emerges. In this section, we discuss some important patterns, and in Section 1.3, we formalize the idea of missingness mechanisms.

Let $Y = (y_{ij})$ denote an $(n \times K)$ rectangular data set without missing values, that is, a complete data set, with i th row $y_i = (y_{i1}, \dots, y_{iK})$, where y_{ij} is the value of variable Y_j for unit i . With missing data, define the *missingness indicator matrix*

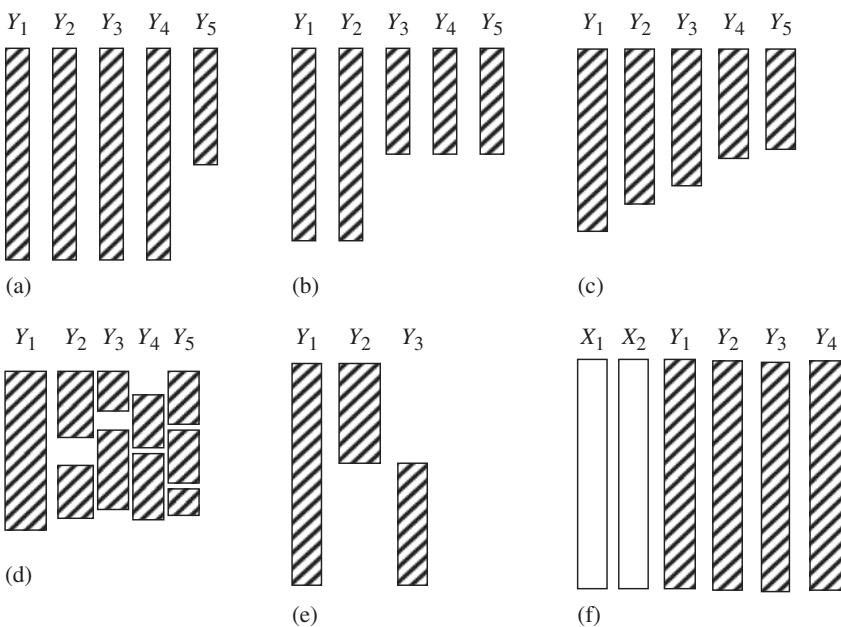


Figure 1.1 Examples of missingness patterns; rows correspond to units and columns to variables. (a) Univariate nonresponse, (b) multivariate with two patterns, (c) monotone, (d) general, (e) file matching, and (f) factor analysis, with two factors and four measured variables.

$M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. The matrix M then defines the pattern of missing data. (A common alternative notation to M is the response indicator matrix R with entries $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ if y_{ij} is missing; to avoid confusion, we use the M notation throughout this book.) It is also possible, and may be useful, to allow different missing data codes that indicate different types of missing data; for example $m_{ij} = 1$ for unit nonresponse due to noncontact, $m_{ij} = 2$ for unit nonresponse due to refusal to participate, and $m_{ij} = 3$ for refusal to answer a particular item. Or in a clinical trial, $m_{ij} = 1$ indicates missing data because the i th participant is terminated, $m_{ij} = 2$ for missing data arising from the i th participant going off a treatment because of side effects, and $m_{ij} = 3$ for missing data because the i th participant can no longer be contacted. Information contained in these codes can be used in the statistical analysis. We usually focus on the simpler situation where m_{ij} is binary.

Figure 1.1 shows some examples of missingness patterns. Some methods for handling missing data apply to any missingness pattern, whereas other methods are restricted to a special pattern.

Example 1.4 *Univariate Missing Data.* Figure 1.1a illustrates *univariate* missing data, where missingness is confined to a single variable. The first

incomplete data problem to receive systematic attention in the statistics literature has the pattern of Figure 1.1a, namely, the problem of missing data in designed experiments. In the context of agricultural trials, the units are experimental plots, and the situation is often called the missing plot problem. Interest is in the relationship between a dependent variable Y_K , such as yield of crop, on a set of factors Y_1, \dots, Y_{K-1} , such as variety, type of fertilizer, and temperature, all of which are intended to be fully observed (in the figure, $K = 5$). Often, a balanced experimental design is chosen that yields orthogonal factors when no values of Y_K are missing, and hence, a simple analysis. Sometimes, however, the outcomes for some of the experimental units are missing (for example because of lack of germination of a seed, or because the data were incorrectly recorded). The result is the missingness pattern with Y_K incomplete and Y_1, \dots, Y_{K-1} fully observed. Missing data techniques can be used to fill in the missing values of Y_K , to retain the balance in the original experimental design. Historically important methods, reviewed in Chapter 2, were motivated by computational simplicity and hence are less important in our era of high-speed computers, but they can still be useful in high-dimensional problems.

Example 1.5 Unit and Item Nonresponse in Surveys. Another common pattern is obtained when the single incomplete variable Y_K in Figure 1.1a is replaced by a set of variables Y_{J+1}, \dots, Y_K , all observed or missing on the same set of units (see Figure 1.1b, where $K = 5$ and $J = 2$). An example of this pattern is unit nonresponse in sample surveys, where a questionnaire is administered, and a subset of sampled individuals does not complete the questionnaire because of noncontact, refusal, or some other reason. In that case, the survey items are the incomplete variables, and the fully observed variables consist of survey design variables measured for both respondents and nonrespondents, such as household location, or characteristics measured in a listing procedure prior to the survey. Common techniques for addressing unit nonresponse in surveys are discussed in Chapter 3. Survey practitioners call missing values on particular items in the questionnaire *item nonresponse*. These missing values typically have a haphazard pattern, such as that in Figure 1.1d. Item nonresponse in surveys is typically handled by imputation methods as discussed in Chapter 4, although the methods discussed in Part II of the book are also appropriate and relevant. For other discussions of missing data in the survey context, see Madow and Olkin (1983), Madow et al. (1983a,b), Rubin (1987a), and Groves et al. (2002).

Example 1.6 Attrition in Longitudinal Studies. Longitudinal studies collect information on a set of units (subjects) repeatedly over time. A common missing data problem is attrition, where units drop out prior to the end of the study and do not return. For example, in panel surveys, members of the panel may drop out before the end of the study because they move to locations that are inaccessible to the researchers. Or in a clinical trial, some units may drop

out for unknown reasons, possibly side effects of drugs, or curing of disease. The pattern of attrition is an example of *monotone* missing data, where the variables can be arranged so that all Y_{j+1}, \dots, Y_K are missing for all units where Y_j is missing, for all $j = 1, \dots, K-1$ (see Figure 1.1c for $K=5$). Methods for handling monotone missing data can be easier to apply than methods for general patterns, as we shall see in Chapter 7 and elsewhere.

In practice, the pattern of missing data is rarely monotone but is often close to monotone. Consider for example the data pattern in Table 1.3, which was obtained from the results of a panel study of students in 10 Illinois schools, analyzed by Marini et al. (1980). The first block of variables was recorded for all individuals at the start of the study, and hence is completely observed. The second block consists of variables measured for all respondents in the follow-up study, 15 years later. Of all respondents to the original survey, 79% responded to the follow-up, and thus the subset of variables in block 2 is regarded as 79% observed. Block 1 variables are consequently *more observed* than block 2 variables.

The data for the 15-year follow-up survey were collected in several phases, and for economic reasons, the group of variables forming the third block was recorded for a subset of those responding to block 2 variables.

Table 1.3 Example 1.6: patterns of missing data across four blocks of variables (0 = observed, 1 = missing)

Pattern	Adolescent variables, block 1	Variables measured for all follow-up respondents, block 2	Variables measured only for initial follow-up respondents, block 3	Parent variables, block 4	Number of units	Percentage of units
A	0	0	0	0	1594	36.6
B	0	0 ^a	0 ^a	1	648	14.9
C	0	0	1	0 ^b	722	16.6
D	0	0 ^a	1	1	469	10.8
E	0	1	1	0 ^b	499	11.5
F	0	1	1	1	420	9.6
Total					4352	100.0

^aUnits falling outside monotone pattern 2 (block 1 more observed than block 4; block 4 more observed than block 2; block 2 more observed than block 3).

^bUnits falling outside monotone pattern 1 (block 1 more observed than block 2; block 2 more observed than block 3; block 3 more observed than block 4).

Source: Marini et al. (1980). Reproduced with permission of Sage Publication, Inc.

Thus, block 2 variables are more observed than block 3 variables. Blocks 1–3 form a monotone pattern of missing data. The fourth block of variables consists of a small number of items measured by a questionnaire mailed to the parents of all students in the original adolescent sample. Of these parents, 65% responded. The four blocks of variables do not form a monotone pattern. However, by sacrificing a relatively small amount of data, monotone patterns can be obtained. The authors analyzed two monotone data sets. First, the values of block 4 variables for patterns C and E (marked with the letter b) are omitted, leaving a monotone pattern with block 1 more observed than block 2, which is more observed than block 3, which is more observed than block 4. Second, the values of block 2 variables for patterns B and D and the values of block 3 variables for pattern B (marked with the letter a) are omitted, leaving a monotone pattern with block 1 more observed than block 4, which is more observed than block 2, which is more observed than block 3. In other examples (such as the data in Table 1.2, discussed in Example 1.6), the creation of a monotone pattern involves the loss of a substantial amount of data.

Example 1.7 *The File-Matching Problem, with Two Sets of Variables Never Jointly Observed.* With large amounts of missing data, it is possible that some variables are never observed together. When this happens, it is important to be aware of the problem because it implies that some parameters measuring the association between these variables are not estimable from the data alone and attempts to estimate them may yield computational problems or misleading results. Figure 1.1e illustrates an extreme version of this problem that arises in the context of combining data from two sources. In this pattern, Y_1 represents a set of variables that is common to both data sources and fully observed, Y_2 a set of variables observed from the first data source but not the second, and Y_3 a set of variables observed from the second data source but not the first. Clearly, there is no information in this data pattern about the partial associations of Y_2 and Y_3 given Y_1 ; in practice, analyses of data with this pattern typically make the strong assumption that these partial associations are zero. This pattern is discussed further in Section 7.5.

Example 1.8 *Patterns with Latent Variables That Are Never Observed.* It can be useful to regard certain problems involving unobserved “latent” variables as missing data problems where the latent variables are completely missing and then apply ideas from missing data theory to estimate the parameters. Consider, for example, Figure 1.1f, where $X = (X_1, X_2)$ represents two latent variables that are completely missing, and $Y = (Y_1, Y_2, Y_3, Y_4)$ is a set of variables that are fully observed. Factor analysis can be viewed as an analysis of the multivariate regression of Y on X for this pattern – that is, a pattern with none of the regressor variables observed! Clearly, some assumptions are needed. Standard forms of factor analysis assume the conditional independence of the components of Y given X . Estimation can be achieved by treating the factors X as

fully missing data. If values of Y are also missing according to a haphazard pattern, then methods of estimation can be developed that treat both X and the unobserved values of Y as missing. This example is examined in more detail in Section 11.3.

Example 1.9 *Missing Data in Clinical Trials.* Clinical trials comparing different treatments on participants often involve repeated measures of the outcome over time and hence can have problems of missing data. One common source of missingness arises when participants no longer take their assigned treatments because of ineffectiveness, side effects, or other reasons. Meinert (1980) distinguished this form of *treatment discontinuation* (which Meinert called treatment dropout) from *analysis dropout*, which arises when outcome measures are not recorded for participants. Treatment discontinuation differs from analysis dropout because outcome measures of individuals might continue to be recorded after individuals discontinue treatment. On the other hand, analysis dropouts may occur when individuals remain on treatment, for example if a clinic visit is missed. The definition of missing data in this book relates to analysis dropout rather than treatment discontinuation, but the latter can be handled using the methods in this book, with the aid of the potential outcomes framework introduced in Example 1.2.

1.3 Mechanisms That Lead to Missing Data

In Section 1.2, we focused on various patterns of missing data. A different issue concerns the mechanisms that lead to missingness, and in particular, the question of whether the fact that variables are missing is related to the underlying values of the variables in the data set. Missingness mechanisms are crucial because the properties of missing data methods depend very strongly on the nature of the dependencies in these mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976a), through the simple device of treating the missingness indicators as random variables and assigning them a distribution. We now give an intuitive overview of this theory, which is treated more precisely in Chapter 6, where the concepts introduced here are more formally defined.

Define the complete data matrix $Y = (y_{ij})$ and the missingness indicator matrix $M = (m_{ij})$ as in Section 1.2. Assume for simplicity that the rows (y_i, m_i) are independent and identically distributed over i . The missingness mechanism is characterized by the conditional distribution of m_i given y_i , say $f_{M|Y}(m_i | y_i, \phi)$, where ϕ denotes unknown parameters. If missingness does not depend on the values of the data, missing or observed, that is, if for all i and any distinct values y_i, y_i^* in the sample space of Y ,

$$f_{M|Y}(m_i | y_i, \phi) = f_{M|Y}(m_i | y_i^*, \phi), \quad (1.1)$$

the data are called missing completely at random (MCAR). Let $y_{(0)i}$ denote the components of y_i that are observed for unit i , and $y_{(1)i}$ denote the components of y_i that are missing for unit i . A less restrictive assumption than MCAR is that missingness depends on y_i only through the observed components $y_{(0)i}$, that is if for all i and any distinct values $(y_{(1)i}, y_{(1)i}^*)$ of the missing components in the sample space of $y_{(1)i}$,

$$f_{M|Y}(m_i | y_{(0)i}, y_{(1)i}, \phi) = f_{M|Y}(m_i | y_{(0)i}, y_{(1)i}^*, \phi). \quad (1.2)$$

The missingness mechanism is then called missing at random (MAR). The mechanism is called missing not at random¹ (MNAR) if the distribution of m_i depends on the missing components of y_i , that is Eq. (1.2) does not hold for some i and some values $(y_{(1)i}, y_{(1)i}^*)$ of the missing components.

As discussed in Chapter 6, MAR is a sufficient condition for pure likelihood and Bayesian inferences to be the valid without modeling the missingness mechanism. For frequentist inferences, the corresponding sufficient condition requires Eq. (1.2) to hold, not only for the observed missingness pattern m_i but also for other patterns that could arise in repeated sampling similarly, the condition that MCAR holds in repeated sampling as well as the sample observed is called missing always completely at random (MACAR).

In the models considered in this book, m_i and y_i are usually assigned a joint distribution, that is both M and Y are treated as random variables. In that case, we show later that if the data are MAR, the predictive distribution of the missing values given the observed values for each unit is independent of pattern. This predictive distribution is then the basis for imputation methods, as discussed in Chapters 4 and 10, and the MAR assumption Eq. (1.2) allows this predictive distribution to be estimated from the observed data.

The simplest data structure with missingness is a univariate random sample where y_i and m_i are both scalar variables. Then we have

$$p(Y = y, M = m | \theta, \phi) = \prod_{i=1}^n f_Y(y_i | \theta) \prod_{i=1}^n f_{M|Y}(m_i | y_i, \phi), \quad (1.3)$$

where $f_Y(y_i | \theta)$ denotes the density of y_i indexed by unknown parameters θ , and $f_{M|Y}(m_i | y_i, \phi)$ is the density of a Bernoulli distribution for the binary missingness indicator m_i , with probability $\Pr(m_i = 1 | y_i, \phi)$ that y_i is missing. If missingness is independent of Y , that is if $\Pr(m_i = 1 | y_i, \phi) = \phi$, a constant that does not depend on y_i , then the missingness mechanism is MCAR (or in this situation, equivalently MAR). If the mechanism depends on y_i , then the mechanism is MNAR because it depends on values of y_i , some of which are missing.

In this simple situation, let r denote the number of responding units with $m_i = 0$, so that the missing values lead to a reduction of sample size from n to

r . We might contemplate carrying out the same analyses on the reduced sample of size r as we intended for the size n sample. For example, if we assume that the values are roughly normally distributed, we might estimate the population mean by the sample mean of the responding units with estimated standard error s/\sqrt{r} , where s is the sample standard deviation of the responding units. This strategy is valid if the mechanism is MCAR because then the observed units are a random subsample of all the units. However, if the data are MNAR, then the analysis based on the responding subsample is generally biased for the parameters of the distribution of Y , including the mean.

Example 1.10 Artificially Created Missing Data in a Univariate Normal Sample. The data in Figure 1.2 provide a concrete illustration of this situation. Figure 1.2a presents a stem and leaf plot (i.e., a histogram with individual values retained) of $n = 100$ standard normal deviates. Under normality, the population mean (zero) for this sample is estimated by the sample mean, which here has the value -0.03 . Figure 1.2b presents a subsample of data obtained from the original sample in Figure 1.2a by deleting units by the MCAR mechanism:

$$\Pr(m_i = 1 | y_i, \phi) = 0.5 \quad \text{for all } y_i, \quad (1.4)$$

that is independently with probability 0.5. The resulting sample of size $r = 52$ is a random subsample of the original values whose sample mean, -0.11 , estimates the population mean of Y without bias.

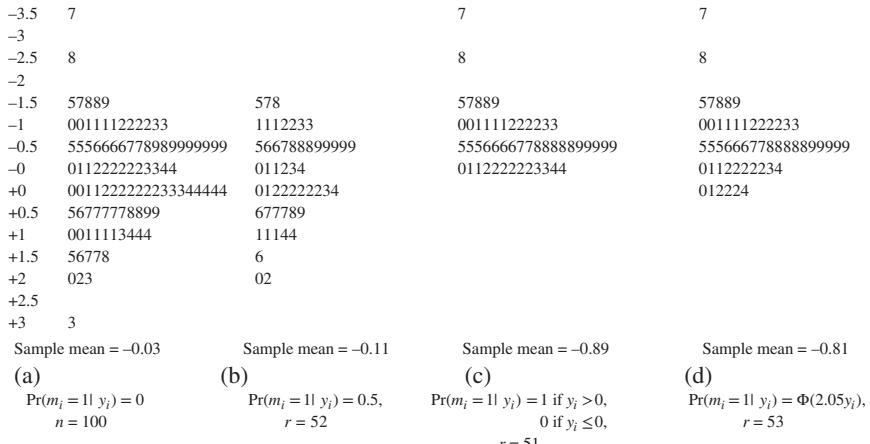


Figure 1.2 Example 1.10: stem and leaf displays of distribution of standard normal sample with (a) no missing data, (b) MCAR missing data, (c) pure censoring, and (d) stochastic censoring.

Figure 1.2c,d illustrates MNAR mechanisms. In Figure 1.2c, negative values from the original sample have been retained and positive values have been deleted, that is

$$\Pr(m_i = 1 \mid y_i, \phi) = \begin{cases} 1, & \text{if } y_i > 0, \\ 0, & \text{if } y_i \leq 0. \end{cases} \quad (1.5)$$

This mechanism is clearly MNAR, and the standard complete data analysis that ignores the missingness mechanism is biased. In particular, the sample mean, -0.89 , clearly underestimates the population mean of Y . The mechanism (1.5) is a form of censoring, with observed values *censored from above*, or *right censored*, at the value zero.

The data in Figure 1.2d are the respondents from the original sample with missingness mechanism:

$$\Pr(m_i = 1 \mid y_i, \phi) = \Phi(2.05y_i), \quad (1.6)$$

where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function. The probability of being missing increases as y_i increases, and thus, most of the observed values are negative. The missingness mechanism is again MNAR, and the sample mean, -0.81 in the example, again systematically underestimates the population mean. The mechanism (1.6) is a form of *stochastic censoring*.

Now suppose that we are faced with an incomplete sample as in Figure 1.2c, and we wish to estimate the population mean. If the missingness mechanism is *known*, then methods are available that correct for the bias of the sample mean, as discussed in Section 15.2. If the missingness mechanism is *unknown*, the problem is much more difficult. The principal evidence that the missingness mechanism is not MAR lies in the fact that the observed values are asymmetrically distributed, which may contradict the assumption that the original data have a (symmetric) normal distribution. If we are confident that the population distribution is symmetric, we can use this information to adjust for bias. On the other hand, if we have little knowledge about the form of the population distribution, we cannot say whether the data are a censored sample from a symmetric distribution or a random subsample from an asymmetric distribution. In the former situation, the sample mean is biased for the population mean; in the latter situation it is not.

Example 1.11 Right-Censored Survival Data. A common practical example of censoring with known censoring points occurs in data where the outcome is the time to an event. Suppose the data have the pattern of Figure 1.1a, Y_1, \dots, Y_{K-1} are fully observed, the variable Y_K measures time to the occurrence of an event (e.g., death of an experimental animal, birth of a child, failure of a light bulb). For some units in the sample, time to occurrence is *right censored*, meaning that the event had not occurred when data collection ended. If the

censoring time is known, then we have the partial information that the value of Y_K exceeds the censoring time. The analysis of the data needs to take account of this information to avoid biased conclusions. Censoring is a particular case of *coarsened* data, and the concept of coarsened at random, which parallels for coarsened data the concept of MAR for simple missing data, is discussed in Section 6.4.

Example 1.12 Historical Heights. Wachter and Trussell (1982) present an interesting illustration of stochastic censoring with unknown censoring points, involving the estimation of heights. The distribution of heights in historical populations is of considerable interest in the biomedical and social sciences because of the information it provides about nutrition, and hence indirectly about living standards. Most of the recorded information concerns the heights of recruits for the armed services. The samples are subject to censoring because minimal height standards were often in place but were enforced with varying strictness, depending on the demand for and supply of recruits. Thus, a typical observed distribution of heights might take the form of the unshaded histogram in Figure 1.3. The shaded area in the figure represents the heights of men excluded from the recruit sample, and is depicted under the assumption that heights are normally distributed in the original uncensored population. Wachter and Trussell discuss methods for estimating the mean and variance of the uncensored distribution under this crucial normal assumption. In this example, there is considerable external evidence that heights in unrestricted populations *are* nearly normal; therefore, the inferences from the stochastically censored data under the assumption of normality have some justification. In many other problems involving missing data, such information is not

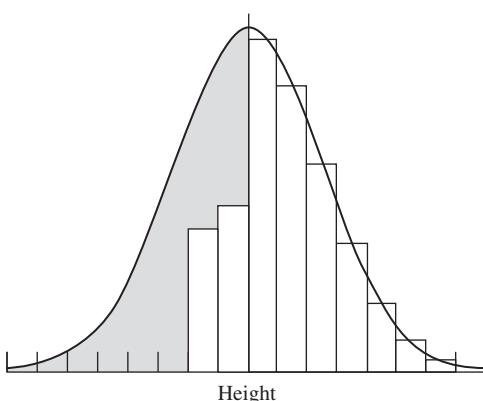


Figure 1.3 Example 1.12: observed and population distributions of historical heights. Population distribution is normal, observed distribution is represented by the histogram, and the shaded area represents missing data.

available or is highly tenuous in nature. As discussed in Chapter 15, the sensitivity of answers from an incomplete sample to critical assumptions is a basic problem in the analysis of data with unknown missingness mechanisms, such as can occur in survey data with nonresponse.

Example 1.13 *MAR for Univariate Missing Data.* Suppose Y is a variable subject to missing values, M is the missingness indicator for Y , and we add a set of fully observed variables X . The data consist of n units, where r units, say $i = 1, \dots, r$ have X and Y observed ($m_i = 0$), and $n - r$ units $i = r + 1, \dots, n$ have X observed and Y missing ($m_i = 1$). We assume that for $i = 1, \dots, n$, (y_i, m_i) are independent given x_i . The MAR assumption Eq. (1.2) about the missingness mechanism for this data structure is

$$f_{M|Y}(m_i | x_i, y_i, \phi) = f_{M|Y}(m_i | x_i, \phi),$$

because missingness of Y cannot depend on Y , as Y is missing for units $i = r + 1, \dots, n$. This implies that M and Y are independent given X , and the conditional distribution of Y given X and M does not depend on M , that is

$$f_{Y|M}(y_i | x_i, m_i = 1, \theta, \phi) = f_{Y|M}(y_i | x_i, m_i = 0, \theta, \phi).$$

This assumption implies that the conditional distribution of Y given X can be estimated for units with Y observed ($M = 0$), and then used to predict the missing values of Y for units with Y missing ($M = 1$). This approach, and extensions for more general missingness patterns, underlies many of the methods in the first 14 chapters of this book. Chapter 15 relaxes the underpinning MAR assumption on the missingness mechanism.

Example 1.14 *Missing Data by Design: Double and Matrix Sampling.* Suppose the data consist of a set of variables (Y_1, \dots, Y_K) , where the first $J < K$ variables Y_1, \dots, Y_J are measured on a sample of size n , and the remaining Y_{J+1}, \dots, Y_K are measured on a subsample of units of size $r < n$. The resulting data have the pattern of Figure 1.1b when $J = 2$ and $K = 5$. A special case arises when $K = J + 1$, Y_{J+1} is a variable of interest but is expensive to measure, and Y_1, \dots, Y_J include inexpensive surrogate measures for Y_{J+1} as well as other variables. The design yields the pattern of Figure 1.1a, with Y_1, \dots, Y_J recorded for the full sample of size n and Y_{J+1} recorded for the subsample of size r . One possible analysis approach is to impute missing values of Y_{J+1} based on the regression of Y_{J+1} on Y_1, \dots, Y_J , computed from the r complete units. This approach is discussed in more detail in Chapter 7.

When the subsample is a probability sample of the original sample, this design is known as double sampling, and the data structure can be treated as a missing data problem, with the values of Y_{J+1}, \dots, Y_K for units not in the subsample being missing. When the subsample is randomly chosen, the missing

data are MCAR, with $\Pr(m_i = 1 | y_{i1}, \dots, y_{iK}) = r/n$. In some situations, it is advantageous to subsample with a rate ϕ that depends on the values of Y_1, \dots, Y_J , as for example where “more interesting” units are subsampled for additional measurements with higher probability. The data are then MAR, with

$$\Pr(m_i = 1 | y_{i1}, \dots, y_{iK}; \phi) = \Pr(m_i = 1 | y_{i1}, \dots, y_{ij}; \phi),$$

where missingness depends on the fully observed variables Y_1, \dots, Y_J but does not depend on the incomplete variables Y_{J+1}, \dots, Y_K . More generally, matrix sampling designs (Mislevy et al. 1992; Raghunathan and Grizzle 1995) partition the set of sampled units into subsamples, and then administer different subsets of the survey variables to each subsample. This is an important approach to limiting respondent burden, and the resulting data structure can be usefully viewed as a missing data problem. An important example is the design for the National Assessment of Educational Progress (2016).

Example 1.15 Measurement Error as a Missing Data Problem. Many studies in epidemiology and elsewhere involve variables recorded with measurement error, which distorts inferences. Specifically, in regression, coefficients of predictor variables subjected to measurement error are typically attenuated, and treatment effects are potentially estimated with bias when variables subject to measurement error are included as covariates (Fuller 1987; Carroll et al. 2006). Adjustments to correct these biases, however, are rarely applied in epidemiological studies (Jurek et al. 2006). A useful approach to measurement error treats the true values of a variable subject to measurement error as missing data (Cole et al. 2006; Guo and Little 2011; Guo et al. 2011).

Information about measurement error is often obtained in a calibration experiment such as a bioassay, where samples with known true values of the variable, say X , are analyzed by the measuring instrument, yield values W subject to measurement error. The regression of the measured values W on the true values X is estimated, yielding a calibration curve (Higgins et al. 1998). The true values of future measurements are then estimated from this curve and treated as the true values in the main analysis. Simulations have shown that this approach yields substantially biased regression estimates when the measurement error is substantial (Freedman et al. 2008; Guo et al. 2011). Better statistical methods treat this as a missing-data problem.

Figure 1.4 displays patterns of missing data on four variables X, W, Y, Z from a main sample and a calibration sample, for two designs, (a) internal calibration and (b) external calibration (Guo and Little 2011; Guo et al. 2011). In both designs, X is the true variable absent measurement error, W is the proxy for X subject to measurement error, and Y and Z are other variables, assumed to be measured without error. The main analysis concerns the regression of Y on

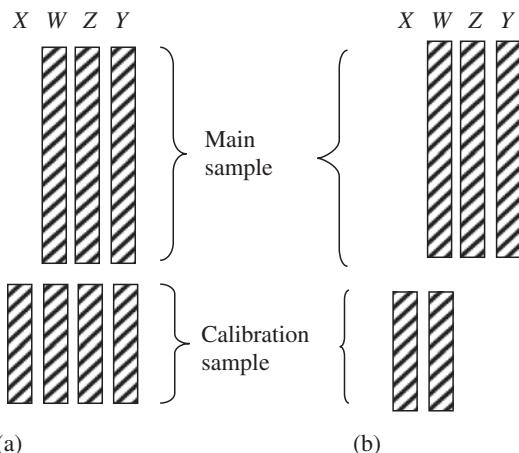


Figure 1.4 Example 1.15: missingness pattern corresponding to main sample data where a variable is measured with error, and a calibration sample where true and measured values of the variable are recorded. X = true covariate, W = measured covariate, Z = other covariates, Y = outcomes. (a) Internal calibration design and (b) external calibration design.

X and Z , and for both designs, the main sample data are a random sample of values of Y , W , and Z , where W is the proxy for X .

Information relating W and X is gained from the calibration sample. In the internal calibration design in Figure 1.4a, all four variables X , W , Y , Z are measured in the calibration sample. The resulting pattern is analogous to double sampling in Example 1.14, and if the calibration sample is a random subsample from the main study, the missingness mechanism is MCAR.

In the external calibration design, only X and W are measured in the calibration sample, yielding the sparser pattern of data shown in Figure 1.4b. This would typically be the case when calibration is carried out independently of the main study, for example by an assay manufacturer. Because the calibration sample is not a random subsample from the main study, in this case the mechanism is typically not MCAR and may not be MAR either. Note also that with external calibration there is no information about parameters for the association between X and (Z, Y) given W ; therefore, assumptions are needed to estimate these parameters.

Treating measurement error as a missing data problem, as in these figures, leads to effective analysis strategies for adjusting for measurement error, as discussed in Section 11.7.

Example 1.16 *Missing Data by Design: Disclosure Limitation.* Another form of designed missing data arises in disclosure limitation. An important feature of surveys is that the identity of the respondent is usually confidential, and information is only used for statistical purposes. The data collector may not wish to

release all the information in a public-use microdata file because the values of some variables may be available from external data bases and may permit a data intruder to discover the identity of some of the respondents. For example, individuals with very high incomes are vulnerable to disclosure if those incomes are made available; therefore, incomes above a particular value are often censored by the data producer. One approach to disclosure limitation is to delete the values of some variables for some respondents and then replace them by artificial multiply-imputed values (Rubin 1993; Little 1993a; Raghunathan et al. 2003; Little et al. 2004; Drechsler 2011). Here, as in Example 1.15, the mechanism for creating missing values is under the control of the data producer and hence usually MCAR or MAR. The effect of replacing true values by synthetic multiply-imputed values can be studied by the data producer because the original values of the variables are available.

Example 1.17 *Income Nonresponse.* In other situations, missing values do not arise by design, but rather because of refusal to answer particular questions. For example, suppose $K = 2$, $Y_1 = \text{age}$ is fully observed, but $Y_2 = \text{income}$ has missing values, yielding the pattern of Figure 1.1a. Unlike in Example 1.14, the missingness mechanism is unknown, and MCAR and MAR become assumptions rather than being known to hold. If the probability that income is missing is the same for all individuals, regardless of their age or income, then the data are MCAR. If the probability that income is missing varies according to the age of the respondent, but does not vary according to the income of those surveyed with the same age, then the data are MAR. If the probability that income is recorded varies according to income for those with the same age, in some partially unknown way, then the data are MNAR. This latter situation is the hardest to deal with analytically, which is unfortunate because it may be the most likely.

When missing data are not under the control of the data collector, the MAR assumption is made more plausible by collecting data Y_1, \dots, Y_{K-1} on respondents and nonrespondents that are predictive both of Y_K and the probability of missingness. Including these data in the analysis then typically replaces the relevant association from that between M and Y_K to that between M and Y_K given Y_1, \dots, Y_{K-1} and thereby helps to make the MAR assumption more plausible as an approximation to reality.

The significance of these assumptions about the missingness mechanism depends somewhat on the objective of the analysis. For example, if interest lies in the marginal distribution of Y_1, \dots, Y_{K-1} , then the data on Y_K , and the mechanism leading to missing values of Y_K , are usually irrelevant (“usually” because one can construct examples where this is not the case, but such examples are often of theoretical rather than practical importance). If interest lies in the conditional distribution of Y_K given Y_1, \dots, Y_{K-1} , as, for example when we are studying how the distribution of income varies according to age, and age is not missing, then the analysis based on the completely recorded units is satisfactory

when the data are MAR. On the other hand, if interest is in the marginal distribution of Y_K (for example summary measures such as the mean of Y_K), then an analysis based on the completely recorded units is generally biased unless the data are MCAR. With complete data on Y_1, \dots, Y_{K-1} and Y_K , the data on Y_1, \dots, Y_{K-1} are typically not useful for estimating the mean of Y_K ; however, when some values of Y_K are missing, the data on Y_1, \dots, Y_{K-1} are useful for this purpose, both in increasing the efficiency with which the mean of Y_K is estimated and in reducing the effects of bias when the data are not MCAR. These points will be examined in more detail in subsequent chapters.

Example 1.18 Mechanisms of Attrition in Longitudinal Data (Example 1.6 Continued). For the monotone pattern of attrition in longitudinal data (Figure 1.1c for $K = 5$), the notation can be simplified by defining a single missing indicator M that now takes the value j if Y_1, \dots, Y_{j-1} are observed and Y_j, \dots, Y_K are missing (that is, dropout occurs between times $(j - 1)$ and j , and M takes the value $K + 1$ for complete units). The missing data (dropout, attrition) mechanism is then MCAR if

$$\Pr(m_i = j \mid y_{i1}, \dots, y_{iK}, \phi) = \phi \quad \text{for all } y_{i1}, \dots, y_{iK},$$

which is a strong assumption and typically contradicted by observed differences in the distributions of background variables across the missingness patterns. Data are MAR if missingness depends on values recorded prior to drop out, but not on values after dropout, that is

$$\Pr(m_i = j \mid y_{i1}, \dots, y_{iK}, \phi) = \Pr(m_i = j \mid y_{i1}, \dots, y_{ij-1}, \phi) \quad \text{for all } y_{i1}, \dots, y_{iK}.$$

Murray and Findlay (1988) provide an instructive example of MAR for longitudinal data from a study of hypertensive drugs, where the outcome was diastolic blood pressure. By definition in the protocol, the unit became a treatment dropout when the observed diastolic blood pressure became too high. By protocol, they were no longer included in the study, and because follow-up measures were not recorded, they were also analysis dropouts, as defined in Example 1.6. This mechanism is not MCAR because it depends on the values of blood pressure. But because blood pressure at the time of dropout was observed before the unit dropped out, the mechanism is MAR: dropout only depends on the observed part of Y . This is a particularly extreme form of MAR mechanism because the distribution of blood pressures at any particular time is sharply differentiated for individuals who drop out at that time vs. those who stay in the study, and using information from the individuals in the study to predict values for dropouts therefore involves extrapolation.

Example 1.19 MAR for a General Bivariate Pattern. The most general missing data pattern for two variables has four kinds of units: complete units, units with only Y_1 observed, units with only Y_2 observed, and units with both

variables missing. If we assume independence of m_i given (y_{i1}, y_{i2}) across units, so that the pattern m_i for unit i depends only on outcomes y_{i1}, y_{i2} for that unit, we can write

$$\Pr(m_{i1} = r, m_{i2} = s \mid y_{i1}, y_{i2}, \phi) = g_{rs}(y_{i1}, y_{i2}, \phi), \quad r, s \in \{0, 1\},$$

where $g_{00}(y_{i1}, y_{i2}, \phi) + g_{10}(y_{i1}, y_{i2}, \phi) + g_{01}(y_{i1}, y_{i2}, \phi) + g_{11}(y_{i1}, y_{i2}, \phi) = 1$. The MAR assumption then implies that $g_{10}(y_{i1}, y_{i2}, \phi) = g_{10}(y_{i2}, \phi)$, because for this pattern y_{i2} is observed and y_{i1} is missing. Applying a similar logic to the other patterns, MAR implies that

$$\begin{aligned} g_{11}(y_{i1}, y_{i2}, \phi) &= g_{11}(\phi), \\ g_{10}(y_{i1}, y_{i2}, \phi) &= g_{10}(y_{i2}, \phi), \\ g_{01}(y_{i1}, y_{i2}, \phi) &= g_{01}(y_{i1}, \phi), \\ g_{00}(y_{i1}, y_{i2}, \phi) &= 1 - g_{10}(y_{i2}, \phi) - g_{01}(y_{i1}, \phi) - g_{11}(\phi), \end{aligned}$$

the last expression arising because the probabilities have to sum to one. This set of assumptions seems unrealistic, in that it requires that missingness of Y_1 depends on Y_2 and missingness of Y_2 depends on Y_1 . A more natural mechanism is that missingness of Y_j depends on Y_j and missingness of Y_1 and missingness of Y_2 are independent, given the data. This yields

$$\begin{aligned} g_{11}(y_{i1}, y_{i2}, \phi) &= g_{1+}(y_{i1}, \phi)g_{+1}(y_{i2}, \phi), \\ g_{10}(y_{i1}, y_{i2}, \phi) &= g_{1+}(y_{i1}, \phi)(1 - g_{+1}(y_{i2}, \phi)), \\ g_{01}(y_{i1}, y_{i2}, \phi) &= (1 - g_{1+}(y_{i1}, \phi))g_{+1}(y_{i2}, \phi), \\ g_{00}(y_{i1}, y_{i2}, \phi) &= (1 - g_{1+}(y_{i1}, \phi))(1 - g_{+1}(y_{i2}, \phi)). \end{aligned}$$

This mechanism is MNAR. The MAR assumption, although sometimes unrealistic, can be a better approximation to reality than the MCAR assumption, which assumes that all four probabilities are unrelated to the outcomes. For instance, in some empirical settings, the MAR assumption has been found to yield more accurate predictions of the missing values than methods based on the “more natural” MNAR mechanism above – see Example 15.18 and Rubin et al. (1996). For a latent variable explanation, see Mealli and Rubin (2015).

1.4 A Taxonomy of Missing Data Methods

Missing data methods proposed in the literature can be usefully grouped into the following categories, which are not mutually exclusive:

1. *Procedures based on completely recorded units*: When some variables are not recorded for some of the units, a simple expedient, mentioned in

Section 1.1, is simply to discard incompletely recorded units and to analyze only the units with complete data (e.g., Nie et al. 1975). This strategy, which we call complete-case analysis, is discussed in Chapter 3. It is easy to implement and may be satisfactory with small amounts of missing data. It can lead to serious biases, however, and it is not usually very efficient, especially when drawing inferences for subpopulations.

2. *Weighting procedures:* Randomization inferences from sample survey data without nonresponse commonly weight sampled units by their *design weights*, which are inversely proportional to their known probabilities of selection. For example, let y_i be the value of a variable Y for unit i in the population. Then, the population mean is often estimated by the Horvitz and Thompson (1952) estimator $\bar{y}_{\text{HT}} = \sum_{i=1}^n \pi_i^{-1} y_i / N$, or the Hajek (1971) estimator

$$\bar{y}_{\text{HK}} = \frac{\sum_{i=1}^n \pi_i^{-1} y_i}{\sum_{i=1}^n \pi_i}, \quad (1.7)$$

where the sums are over the n sampled units, N is the population size, and π_i is the known probability of inclusion in the sample for unit i . Weighting procedures for nonresponse modify the weights in an attempt to adjust for nonresponse as if it were part of the sample design. The resultant estimator (1.7) is replaced by

$$\frac{\sum_{i=1}^n (\pi_i \hat{p}_i)^{-1} y_i}{\sum_{i=1}^n (\pi_i \hat{p}_i)},$$

where the sums are now over sampled units that respond, and \hat{p}_i is an estimate of the probability of response for unit i , often the proportion of responding units in the subclass of the sample in which unit i falls. Weighting methods are described further in Chapter 3.

3. *Imputation:* The missing values are filled in, and the resultant completed data are analyzed by standard methods. Commonly used procedures for imputation include *hot deck* imputation, where units without missingness are used to substitute values; *mean* imputation, where means from units with recorded values are substituted; and *regression* imputation, where the missing variables for a unit are estimated by predicted values from the regression on the known variables for that unit. Principles of imputation are discussed in Chapter 4. For valid inferences to result, modifications to the standard analyses are required to allow for the differing status of the real and the imputed values. Approaches for measuring and incorporating imputation uncertainty are discussed in Chapter 5, including multiple imputation, which reappears in Parts II and III in the context of model-based methods.

4. *Model-based methods:* A broad class of procedures is generated by defining a model for the complete data and basing inferences on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood. Advantages of this approach are flexibility; the avoidance of ad hoc methods, in that model assumptions underlying the resulting methods can be displayed and evaluated; and the availability of estimates of sampling variance that take into account incompleteness in the data. Model-based procedures are the main focus of this book and are developed in Chapters 6–15, which comprise Parts II and III.

Example 1.20 *Estimating the Mean and Covariance Matrix with Monotone Missingness Pattern.* Many multivariate statistical analyses, including least squares regression, factor analysis, and discriminant analysis, are typically based on an initial reduction of the data to the sample mean vector and sample covariance matrix of the variables. An important practical question is how to estimate these quantities from incomplete data. Early literature, discussed selectively in Chapter 3, proposed ad hoc solutions. A more systematic likelihood-based approach, the focus of Part II, is introduced in Chapter 6 and applied to a variety of situations in the following chapters.

Suppose the data can be arranged in a monotone pattern. A simple approach to estimating the mean and covariance matrix is to restrict the analysis to the units with all variables observed. This method of analysis, however, may discard a considerable amount of data. Also, for many examples, including the data summarized in Table 1.3, the completely observed units are not a random sample of the original sample (i.e., the data are not MCAR), and the resulting estimates may be biased. A more generally successful strategy is to assume that the data have a multivariate normal distribution and to estimate the mean vector and covariance matrix by maximum likelihood. In Chapter 6, we show that for monotone missing data, this task is not as difficult as one might suppose because estimation is simplified by a factorization of the joint distribution of the variables.

In particular, for bivariate monotone missing data with Y_1 fully observed and Y_2 subject to missing values, the joint distribution of Y_1 and Y_2 can be factored into the marginal distribution of Y_1 and the conditional distribution of Y_2 given Y_1 . Under simple assumptions, inference about the marginal distribution of Y_1 can be based on all the units, and inferences about the conditional distribution of Y_2 given Y_1 can be based on the subset of units with both Y_1 and Y_2 observed. The results of these analyses can then be combined to estimate the joint distribution of Y_1 and Y_2 and any parameters of this distribution. Estimation of the conditional distribution of Y_2 given Y_1 is a form of regression analysis, and the strategy of factoring the distribution relates to the idea of *imputing* the missing values of Y_2 by regressing Y_2 on Y_1 and then calculating predictions from the regression equation.

Example 1.21 *Estimating the Mean and Covariance Matrix with General Missingness Patterns.* Many data sets with missing values do not exhibit monotone patterns or convenient close approximations such as that displayed in Table 1.3. Methods for estimating the mean and covariance matrix of a set of variables have also been developed that can be applied to any pattern of missing values. As in Example 1.20, these methods are often based on maximum likelihood estimation, assuming the variables are multivariate normally distributed, and this estimation involves iterative algorithms.

The expectation–maximization (EM) algorithm (Dempster et al. 1977), developed in Chapter 8, is an important general technique for finding maximum likelihood estimates from incomplete data. It is applied to multivariate normal data in Chapter 11. The resulting algorithm is closely related to an iterative version of a method that imputes estimates of the missing values by regression. Thus, even in this complex problem, a link can be established between efficient model-based methods and more traditional ad hoc approaches based on substituting reasonable estimates for missing values. Chapter 11 also presents more esoteric uses of the EM algorithm to handle problems such as variance components models, factor analysis, and time series, which can be viewed as missing data problems for multivariate normal data with specialized parametric structure. Bayesian methods for multivariate normal data with missing values are also described in Chapter 11. Robust methods for continuous data with longer tails than the normal are developed in Chapter 14.

Example 1.22 *Estimation When Some Variables Are Categorical.* The reduction of the data to a vector of means and a covariance matrix is generally not appropriate when some variables are categorical. When all variables are categorical and some have missing values, the data can be arranged as a contingency table with partially classified margins as in Example 1.1. Methods for analyzing such data are discussed in Chapter 12. More generally, Chapter 13 considers multivariate data where some of the variables are continuous and some are categorical. That chapter also considers the estimation of finite mixture distributions from a missing-data perspective.

Example 1.23 *Estimation When the Data May Be Missing Not at Random.* Much of the literature on multivariate incomplete data assumes that the data are MAR or MCAR. Chapter 15 deals explicitly with the situation where the data are MNAR, and models are needed for the missingness mechanism. Because it is rarely feasible to estimate the mechanism with any degree of confidence, the main thrust of these methods is to conduct sensitivity analyses to assess the effect of alternative assumptions about the missingness mechanism.

5. *Hybrid approaches:* Approaches based on estimating equations have been proposed that combine the aspects of modeling and weighting. In particular,

augmented inverse probability weighted generalized estimating equations (Robins et al. 1995; Rotnitzky et al. 1998; Seaman and White 2011) combine predictions from models with weighted residuals, designed to protect against model misspecification. Our general view is that careful modeling to avoid serious forms of misspecification renders this hybrid approach unnecessary.

Problems

- 1.1 Find the monotone pattern for the data of Table 1.1 that involves minimal deletion of observed values. Can you think of better statistical criteria for deleting values than this one?
- 1.2 List methods for handling missing values in an area of statistical application of interest to you, based on experience or relevant literature.
- 1.3 What assumptions about the missingness mechanism are implied by the statistical analyses used in Problem 1.2? Do these assumptions appear realistic?
- 1.4 What impact does the occurrence of missing values have on (a) estimates and (b) tests and confidence intervals for the analyses in Problem 1.2? For example, are estimates consistent for underlying population quantities, and do tests have the stated significance levels?
- 1.5 Let $Y = (y_{ij})$ be the data matrix and let $M = (m_{ij})$ be the corresponding missingness indicator matrix, where $m_{ij} = 1$ indicates missing and $m_{ij} = 0$ indicates observed.
 - (a) Propose situations where two values of m_{ij} are not sufficient. (*Hint:* See Heitjan and Rubin (1990).)
 - (b) Nearly always it is assumed that M is fully observed. Describe a realistic case when it may make sense to regard part of M itself as missing. (*Hint:* Can you think of a situation where the meaning of a “blank” is unclear?)
 - (c) Consider the simple situation where $m_{ij} = 1$ or $m_{ij} = 0$. When attention is focused only on the units that fully respond, the conditional distribution of y_i given $m_i = (0, 0, \dots, 0)$ is being estimated, where y_i and m_i are the i th rows of Y and M , respectively. Propose situations where it makes sense to define the conditional distribution of y_i given other missingness patterns. Propose situations where it makes no sense to define these other distributions.
 - (d) Express the marginal distribution of y_i in terms of the conditional distributions of y_i given the various missingness patterns and their probabilities.

- 1.6** Generate 100 triplets $\{(z_{i1}, z_{i2}, z_{i3}), i = 1, \dots, 100\}$ of independent standard normal (that is, mean 0, variance 1) deviates. From these triplets, create 100 trivariate normal observations $\{(y_{i1}, y_{i2}, u_i), i = 1, \dots, 100\}$ on (Y_1, Y_2, U) as follows:

$$y_{i1} = 1 + z_{i1},$$

$$y_{i2} = 5 + 2 \times z_{i1} + z_{i2},$$

$$u_i = a \times (y_{i1} - 1) + b \times (y_{i2} - 5) + z_{i3},$$

with $a = b = 0$. (We choose other values of a and b later.) The units $\{(y_{i1}, y_{i2}): i = 1, \dots, 100\}$ then have a bivariate normal distribution with means $(1, 5)$, variances $(1, 5)$, and correlation $2/\sqrt{5} = 0.89$. Suppose U is a latent variable that is never observed, Y_1 is fully observed, and Y_2 has missing values. For $i = 1, \dots, 100$, create a missingness indicator m_{i2} for y_{i2} based on the value u_i , as follows:

$$\Pr(m_{i2} = 1 | y_{i1}, y_{i2}, u_i, \phi) = \begin{cases} 1, & \text{if } u_i < 0, \\ 0, & \text{if } u_i \geq 0. \end{cases}$$

In other words, Y_2 is missing when U is less than zero. Because U has mean zero, this mechanism should create missing values of Y_2 for about 50% of the units.

- A.** Display the marginal distributions of Y_1 and Y_2 for complete and incomplete units. (Note that in reality the marginal distribution of Y_2 is not available for missing units.) Is this mechanism MCAR, MAR, or MNAR?
- B.** Conduct a t -test comparing the means of Y_1 for complete and incomplete units. Is there evidence from this test that the data are not (a) MCAR, (b) MAR, and (c) MNAR?
- C.** Repeat parts (A) and (B) with (i) $a = 2, b = 0$ and (ii) $a = 0$ and $b = 2$.

Note

¹ In previous editions, we use the term not missing at random (NMAR), but MNAR seems clearer.

2

Missing Data in Experiments

2.1 Introduction

An important problem, historically, occurs with missing outcome data in controlled experiments. This was arguably the first missing data problem to be systematically treated in a principled way, and its treatment anticipated modern missing data methods, particularly the expectation–maximization (EM) algorithm (Dempster et al. 1977) discussed in Chapter 8.

Controlled experiments are generally carefully designed to allow revealing statistical analyses to be made using straightforward computations. In particular, corresponding to a standard classical experimental design, there is a standard least squares analysis, which yields estimates of parameters, standard errors for contrasts of parameters, and the analysis of variance (ANOVA) table. The estimates, standard errors, and ANOVA table corresponding to most designed experiments are easily computed because of balance in the designs. For example with two factors being studied, the analysis is particularly simple when the same number of units is assigned to each combination of factor levels. Textbooks on experimental design catalog many examples of specialized analyses (Box et al. 1985; Cochran and Cox 1957; Davies 1960; Kempthorne 1952; Winer 1962; Wu and Hamada 2009).

Because the levels of the design factors in an experiment are fixed by the experimenter, missing values, if they occur, do so in the outcome variable, Y , rather than in the design factors, X . Consequently, we restrict attention to missing values in Y , and the data have the pattern of Figure 1.1a with (Y_1, \dots, Y_4) representing the fully observed factors (X) and Y_5 representing the incomplete outcome (Y). With this pattern, and assuming missing at random (MAR) and fixed X , the units with Y missing provide no information for the regression of Y on X , so an analysis of the complete cases is fully efficient. However, the balance present in the original design is destroyed by the missing data. As a result, the proper least squares analysis becomes more complicated computationally. An intuitively attractive approach is to fill in the missing values to restore the

balance and then proceed with the standard analysis. This idea of filling in missing data to take advantage of standard analyses recurs frequently in this text.

The advantages of filling in the missing values in an experiment rather than trying to analyze the actual observed data include the following: (i) It is easier to specify the data structure using the terminology of experimental design (for example as a balanced incomplete block), (ii) it is easier to compute necessary statistical summaries, and (iii) it is easier to interpret the results of analyses because standard displays and summaries can be used. Ideally, we would hope that simple rules could be devised for filling in the missing values in such a way that the resultant complete data analyses would be correct. In fact, much progress toward this goal can be made, especially in this context of classical experiments.

Assuming that the missingness is unrelated to the missing values of the outcome variable (i.e., under MAR), there exist a variety of methods for filling in values that yield correct least-squares estimates of all estimable effect parameters. Furthermore, it is easy to correct the residual (error) mean square, standard errors, and sums of squares that have one degree of freedom. Unfortunately, it is more complicated computationally to provide correct sums of squares with more than one degree of freedom, but it can be done, as shown below.

Methods that fill in one value per missing value strictly apply only to analyses based on one fixed-effect linear model with one error term. Examples involving the fitting of more than one fixed-effect linear model include hierarchical models, which attribute sums of squares to effects in a particular order by fitting a sequence of nested fixed-effect models; split-plot and repeated measures designs, which use different error terms for different effects; and random and mixed-effect models, which treat some parameters as random variables. For analyses employing more than one fixed-effect model, in general a different set of missing values have to be filled in for each fixed-effect model. Early discussions in the context of ordinary least-squares estimation are Anderson (1946) and Jarrett (1978).

2.2 The Exact Least Squares Solution with Complete Data

Let X be an $n \times p$ matrix whose i th row, $x_i = (x_{i1}, \dots, x_{ip})$, provides the values of the fixed factors for the i th unit. For example in a 2×2 design with two units per cell and the levels of the factors labeled 0 and 1, we have

$$X = \begin{pmatrix} 100 \\ 100 \\ 101 \\ 101 \\ 110 \\ 110 \\ 111 \\ 111 \end{pmatrix},$$

where the first column represents the intercept, the second column the first factor, and the third column the second factor. The outcome variable $Y = (y_1, \dots, y_n)^T$ is assumed to satisfy the linear model

$$Y = X\beta + e, \quad (2.1)$$

where $e = (e_1, \dots, e_n)^T$ and the e_i are independent and identically distributed with zero mean and common variance σ^2 ; β is the parameter to be estimated, a $p \times 1$ vector. The least squares estimate of β is

$$\hat{\beta} = (X^T X)^{-1} (X^T Y) \quad (2.2)$$

if $X^T X$ is full rank and is undefined otherwise. If $X^T X$ is full rank, $\hat{\beta}$ is the minimum variance unbiased estimate of β . If the e_i are normally distributed, $\hat{\beta}$ is also the maximum likelihood estimate of β (see Chapter 6), and it is normally distributed with mean β and variance $(X^T X)^{-1} \sigma^2$.

The best (minimum variance) unbiased estimate of σ^2 is

$$s^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-p}, \quad (2.3)$$

where $\hat{y}_i = x_i \hat{\beta}$; if the e_i are normal, then $(n-p)s^2/\sigma^2$ is distributed as χ^2 on $n-p$ degrees of freedom. The best unbiased estimate of the covariance matrix of $(\hat{\beta} - \beta)$ is

$$V = (X^T X)^{-1} \hat{\sigma}^2. \quad (2.4)$$

If the e_i are normal, then $(\hat{\beta}_i - \beta_i)/\sqrt{v_{ii}}$ (where v_{ii} is the i th diagonal element of V) has a t distribution on $n-p$ degrees of freedom; $(\hat{\beta} - \beta)$ has a scaled multivariate t distribution with scale $V^{1/2}$.

Tests of the hypothesis that some set of linear combinations of β are all zero are conducted by calculating the sum of squares attributable to the set of linear combinations. More precisely, suppose C is a $p \times w$ matrix specifying the w linear combinations of β that are to be tested; then the sum of squares attributable to the w linear combinations is

$$SS = (C^T \hat{\beta})^T [C^T (X^T X)^{-1} C]^{-1} (C^T \hat{\beta}). \quad (2.5)$$

The test that $C^T \beta = 0$, a vector of zeros, is made by comparing SS/w to $\hat{\sigma}^2$:

$$F = \frac{SS/w}{\hat{\sigma}^2}. \quad (2.6)$$

If the e_i are normal, then F in (2.6) is proportional to the likelihood ratio test statistic for $C^T \beta = 0$; if in addition $C^T \beta = 0$, then F is distributed as Snedecor's F distribution with w and $n-p$ degrees of freedom because SS/σ^2

and $(n-p)s^2/\sigma^2$ are then independent χ^2 random variables with w and $n-p$ degrees of freedom, respectively. Proofs of the preceding results can be found in standard books on regression analysis, such as Draper and Smith (1981) and Weisberg (1980). Also, as these references point out, careful interpretation of these tests may be required in nonorthogonal designs; for example this test of a collection of A effects in a model with A effects, B effects, and interactive effects addresses A adjusted for B main effects and AB interactions.

Standard experimental designs are chosen to make estimation and testing computationally easy and relatively precise. In particular, the matrix $X^T X$ is usually easy to invert, with the result that $\hat{\beta}$, $\hat{\sigma}^2$, V , and the sum of squares attributable to specific collections of linear combinations of β , called contrasts, such as treatment effects and block effects, are easy to calculate. In fact, these summaries of the data are usually calculated as sums of the y_i values and their squares. This simplicity can be important in experiments with several factors and many parameters to be estimated because then $X^T X$ can be of large dimension. The inversion of large matrices was especially cumbersome before the days of modern computing equipment but still can be troublesome in some computing environments when p is very large.

2.3 The Correct Least Squares Analysis with Missing Data

Suppose that X represents the factors in an experimental design such that if all of Y were observed, the analysis of the data would be conducted using existing standard formulas and computer programs. The question of interest here is how to use these complete-data formulas and computer programs when part of Y is missing.

Assuming that the reason for the occurrence of missing data in Y does not depend on any missing Y values (i.e., under MAR) and that the parameters of the missing data process are distinct from the ANOVA parameters (that is these two sets of parameters lie in disjoint parameter spaces), the incomplete units carry no information for the ANOVA model, and the correct approach simply ignores the rows of X corresponding to missing y_i and conducts the least squares analysis described in Section 2.2 using the complete cases with x_i and y_i observed. There are two potential problems with this approach, one statistical and one computational.

The statistical problem is that the design matrix restricted to the observed units may not be positive definite so that the least squares estimates may not be uniquely estimable from the data. Dodge (1985) provides a detailed discussion of this problem, including procedures for detecting it and for determining which treatment effects remain estimable. We assume in what follows that the design matrix based on the complete units is positive definite. In that case, the equations given in Section 2.2 applied to the r units with y_i observed define

the correct least squares estimates, standard errors, sums of squares, and F tests when faced with missing data. We let $\hat{\beta}_*$, $\hat{\sigma}_*^2$, V_* , and S_* denote the quantities in Eqs. (2.2)–(2.5) calculated from these complete units.

The computational problem is that the specialized formulas and computing routines used with complete Y cannot be used because the original balance is no longer present. The remainder of this chapter describes how to obtain these summaries essentially using only the procedures needed for complete data, which use the special structure in X to simplify computations.

2.4 Filling in Least Squares Estimates

2.4.1 Yates's Method

The classical and standard approach to missing data in ANOVA is due in general to Yates (1933). Yates noted that if the missing values were replaced by their least squares estimates $\hat{y}_i = x_i \hat{\beta}_*$, where $\hat{\beta}_*$ is defined by (2.2) applied to the r rows of (Y, X) that have y_i observed, then the method of least squares applied to the filled-in data yields the correct least squares estimates, $\hat{\beta}_*$. This approach of filling in least squares estimates may seem circular and of little practical help because it appears to require knowledge of $\hat{\beta}_*$ to estimate the missing y_i as $x_i \hat{\beta}_*$ before $\hat{\beta}_*$ can be calculated. It turns out, perhaps surprisingly, that it can be relatively easy to calculate $\hat{y}_i = x_i \hat{\beta}_*$ for the missing y_i before calculating $\hat{\beta}_*$ directly, at least if only a few values are missing.

The rationale for Yates's procedure is that it yields (i) the correct least squares estimates of β , $\hat{\beta}_*$ and (ii) the resultant residual sum of squares equals $(r - p)s_*^2$, so division by $(r - p)$ rather than $(n - p)$ yields the correct least squares estimate s_*^2 of σ^2 . It is quite easy to prove these two facts. For the results in this chapter, a convenient notation lets $i = 1, \dots, m$ index the m missing values and $i = m + 1, \dots, n$ index the $r = n - m$ observed values. Let $\hat{y}_i = x_i \hat{\beta}_*$, $i = 1, \dots, m$ denote the least squares estimates of the m missing values. Complete-data methods applied to the filled-in data minimize the quantity

$$SS(\beta) = \sum_{i=1}^m (\hat{y}_i - x_i \beta)^2 + \sum_{m+1}^n (y_i - x_i \beta)^2$$

with respect to β . By definition, $\beta = \hat{\beta}_*$ minimizes the second summation in $SS(\beta)$ but also by definition, $\beta = \hat{\beta}_*$ minimizes the first summation in $SS(\beta)$, setting it equal to zero. Consequently, with least squares estimates of missing values filled in, $SS(\beta)$ is minimized at $\beta = \hat{\beta}_*$, and $SS(\hat{\beta}_*)$ equals the minimal residual sum of squares over the r observed values of y_i . Hence, the correct least squares estimate of β , $\hat{\beta}_*$ equals the least squares estimate of β found by the complete-data ANOVA program, and the correct least squares

estimate of σ^2 , s_*^2 is found from the complete-data ANOVA estimate of σ^2 , s^2 , by

$$s_*^2 = s^2 \frac{n - p}{r - p}.$$

The analysis of the filled-in data with missing y_i set equal to \hat{y}_i is not perfect: It yields an estimated covariance matrix of $\hat{\beta}$ that is too small and sums of squares attributable to collections of linear combinations of β that are too big, although for small fractions of missing data, these biases are often relatively minor. We now consider methods for calculating the values \hat{y}_i .

2.4.2 Using a Formula for the Missing Values

One approach is to use a formula for the missing values, fill them in, and then proceed. In the first application of this idea, Allan and Wishart (1930) provided formulas for the least squares estimate of one missing value in a randomized block design and of one missing value in a Latin square design. For example, in a randomized block with T treatments and B blocks, the least squares estimate of a missing value in treatment t and block b is

$$\frac{T y_+^{(t)} + B y_+^{(b)} - y_+}{(T - 1)(B - 1)},$$

where $y_+^{(t)}$ and $y_+^{(b)}$ are the sum of the observed values of Y for treatment t and block b , respectively, and y_+ is the sum of all observed values of Y . Wilkinson (1958a) extended this work by giving tables providing formulas for many designs and many patterns of missing values.

2.4.3 Iterating to Find the Missing Values

Hartley (1956) proposed a general noniterative method for estimating one missing value that he suggested should be used iteratively for more than one. The method for one missing value involves substituting three different trial values for the missing value, with the residual sum of squares calculated for each trial value. Because the residual sum of squares is quadratic in one missing value, the minimizing value of the one missing value can then be found. This method, although clever, is not as attractive as alternative methods.

Healy and Westmacott (1956) described a popular iterative technique that is sometimes attributed to Yates and even sometimes to Fisher. With this method, (i) trial values are substituted for all missing values, (ii) the complete data analysis is performed, (iii) predicted values are obtained for the missing values, (iv) these predicted values are substituted for the missing values, (v) a new complete-data analysis is performed, and so on, until the missing values do not

change appreciably, or equivalently, until the residual sum of squares essentially stops decreasing.

We show later in Example 11.4 that the Healy and Westmacott method for estimating β is an example of an EM algorithm, introduced here in Chapter 8, and each iteration decreases the residual sum of squares (or equivalently, increases the likelihood under the corresponding normal linear model). In some cases, convergence can be slow and special acceleration techniques have been suggested (Pearce 1965, p. 111; Preece 1971). Although these can improve the rate of convergence in some examples, they can also destroy the monotone decrease of the residual sum of squares in other examples (Jarrett 1978).

2.4.4 ANCOVA with Missing Value Covariates

A general noniterative method due to Bartlett (1937) is to fill in guesses for the missing values and then perform an analysis of covariance (ANCOVA) with a missing value covariate for *each* missing value. The i th missing value covariate is defined to be the indicator for the i th missing value, that is, zero everywhere except for the i th missing value where it equals one. The coefficient of the i th missing value covariate, when subtracted from the initial guess of the i th missing value, yields the least squares estimate of the i th missing value. Furthermore, the residual mean square and all contrast sums of squares adjusted for the missing value covariates are their correct values. We prove these results in Section 2.5.

Although this method is quite attractive in some ways, it often cannot be directly implemented because specialized ANOVA routines may not have the capability to handle multiple covariates. It turns out, however, that Bartlett's method can be applied using only the existing complete-data ANOVA routine and a routine to invert an $m \times m$ symmetric matrix. The next section proves that Bartlett's method leads to the correct least squares analysis. The subsequent section concerns how to obtain this analysis using only the complete-data ANOVA routine.

2.5 Bartlett's ANCOVA Method

2.5.1 Useful Properties of Bartlett's Method

Bartlett's ANCOVA method has the following useful properties. First, it is noniterative and thus avoids questions of convergence. Second, if there is a singular pattern of missing values (i.e., a pattern such that some parameters are inestimable as when all values under one treatment are missing), the method will warn the user, whereas iterative methods will produce an answer, possibly a quite inappropriate one. A third advantage is, as mentioned earlier,

that the method produces not only the correct estimates and correct residual sum of squares but also correct standard errors, sums of squares, and F tests as well.

2.5.2 Notation

Suppose that each missing y_i is filled in with some initial guess in order to create a complete vector of values for Y . Call the initial guesses \tilde{y}_i , $i = 1, \dots, m$. Also, let Z be the $n \times m$ matrix of m missing value covariates: the first row of Z equals $(1, 0, \dots, 0)$, the m th row equals $(0, \dots, 0, 1)$, and the last r rows of Z equal $(0, 0, \dots, 0)$ because they correspond to observed y_i . The ANCOVA uses both X and Z to predict Y .

Analogous to (2.1), the model for Y is now

$$Y = X\beta + Z\gamma + e, \quad (2.7)$$

where γ is a column vector of m regression coefficients for the missing value covariates. The residual sum of squares to be minimized over (β, γ) is

$$\text{SS}(\beta, \gamma) = \sum_{i=1}^m (\tilde{y}_i - x_i\beta - z_i\gamma)^2 + \sum_{i=m+1}^n (y_i - x_i\beta - z_i\gamma)^2.$$

Because $z_i\gamma = 0$ when y_i is observed and $z_i\gamma = \gamma_i$ when y_i is missing,

$$\text{SS}(\beta, \gamma) = \sum_{i=1}^m (\tilde{y}_i - x_i\beta - \gamma_i)^2 + \sum_{i=m+1}^n (y_i - x_i\beta)^2. \quad (2.8)$$

2.5.3 The ANCOVA Estimates of Parameters and Missing Y -Values

As before, let $\hat{\beta}_*$ equal the correct least squares estimate of β obtained by applying (2.2) to the observed values, that is to the last r rows of (Y, X) ; this minimizes the second summation in (2.8). But with $\beta = \hat{\beta}_*$, setting γ equal to $(\hat{\gamma}_1, \dots, \hat{\gamma}_m)^T$ where

$$\hat{y}_i = \tilde{y}_i - x_i\hat{\beta}_*, \quad i = 1, \dots, m \quad (2.9)$$

minimizes the first summation in (2.8) by making it identically zero, so that

$$\text{SS}(\hat{\beta}_*, \hat{\gamma}) = \sum_{i=m+1}^n (y_i - x_i\hat{\beta}_*)^2. \quad (2.10)$$

Thus, $(\hat{\beta}_*, \hat{\gamma})$ minimizes $\text{SS}(\beta, \gamma)$ and gives the least squares estimate of (β, γ) obtained from the ANCOVA model in (2.7). Equation (2.9) also implies that

the correct least squares estimate of the missing y_i , that is, $\hat{y}_i = \mathbf{x}_i \hat{\beta}_*$, is given by $\tilde{y}_i - \hat{y}_i$ or in words:

Correct least squares predicted value for i th missing value = initial guess
for i th missing value – coefficient of i th missing value covariate. (2.11)

Bartlett's original description of this method set all \tilde{y}_i equal to zero, but setting all \tilde{y}_i equal to the grand mean of all units is computationally more attractive and also yields the correct total sum of squares about the grand mean.

2.5.4 ANCOVA Estimates of the Residual Sums of Squares and the Covariance Matrix of $\hat{\beta}$

Equation (2.10) establishes that the residual sum of squares from the ANCOVA is the correct residual sum of squares; the ANCOVA degrees of freedom corresponding to this residual sum of squares is $n - m - p = r - p$, which is also correct. Consequently, the residual mean square is correct and equal to s_*^2 . If the covariance matrix of $\hat{\beta}_*$ from the ANCOVA equals V_* obtained by applying (2.4) to the r units with y_i observed, then all standard errors, sums of squares, and tests of significance will also be correct. The estimated covariance matrix of $\hat{\beta}_*$ from the ANCOVA is the estimated residual mean square, s_*^2 , times the upper left $p \times p$ submatrix of $((X, Z)^T(X, Z))^{-1}$, say U . Because the estimated residual mean square is correct, we need only show that U^{-1} is the sum of cross products of X for the units with y_i observed. From standard results on matrices,

$$U = [X^T X - (X^T Z)(Z^T Z)^{-1}(Z^T X)]^{-1}. \quad (2.12)$$

By the definition of z_i ,

$$X^T Z = \sum_{i=1}^m x_i^T z_i \quad (2.13)$$

and

$$Z^T Z = \sum_{i=1}^m z_i^T z_i = I_m, \quad (2.14)$$

the $(m \times m)$ identity matrix. From (2.13) and (2.14),

$$(X^T Z)(Z^T Z)^{-1}(Z^T X) = \left(\sum_{i=1}^m x_i^T z_i \right) \left(\sum_{j=1}^m z_j^T x_j \right). \quad (2.15)$$

But

$$z_i z_j^T = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

whence (2.15) equals

$$\sum_{i=1}^m x_i^T x_i,$$

and from (2.12)

$$U = \left(\sum_{i=m+1}^n x_i^T x_i \right)^{-1},$$

so that $s_*^2 U = V_*$, the covariance matrix of $\hat{\beta}_*$ found by ignoring the missing units, as required to complete the proof that Bartlett's ANCOVA produces least squares values for all summaries.

2.6 Least Squares Estimates of Missing Values by ANCOVA Using Only Complete-Data Methods

The preceding theory relating incomplete-data ANOVA and a complete data ANCOVA would be merely of academic interest if the ANCOVA needed special software to implement. We now describe how to implement the missing-value covariate method to calculate least squares estimates of m missing values using only complete-data ANOVA routines and a routine to invert an $m \times m$ symmetric matrix (the sweep operator, described in Section 7.5, can be used for this purpose). In Section 2.7, the analysis is extended to produce correct standard errors and sums of squares for one degree of freedom hypotheses. The argument here will appeal to the ANCOVA results; a direct algebraic proof appears in Rubin (1972).

By ANCOVA theory, the vector $\hat{\gamma}$ can be written as

$$\hat{\gamma} = B^{-1} \rho, \quad (2.16)$$

where B is the $m \times m$ cross-products matrix for the residuals of the m missing value covariates after adjusting for the matrix X , and ρ is the $m \times 1$ vector of cross products of Y and the residuals of the missing-value covariates after adjusting for the X matrix. If B is singular, the pattern of missing data is such that an attempt is being made to estimate inestimable parameters, such as the effect of a treatment when all units exposed to that treatment are missing. The method requires (i) the calculation of B and ρ using the complete-data ANOVA routine, (ii) the inversion of B to obtain $\hat{\gamma}$ from (2.16), and (iii) the calculation of the missing values from (2.11).

To find B and ρ , begin by performing a complete-data ANOVA on the first missing-value covariate, that is, using the first column of Z (which is all zeros except for a one where the first missing value occurs), as the dependent variable rather than Y . The residuals from this analysis for the m missing values comprise the first row of B . Repeat the complete data ANOVA on the j th missing value covariate, $j = 2, \dots, m$ (which is all zeros except for a one where the j th missing value occurs), and let the j th row of B equal the residuals for the m missing values from this analysis. The vector ρ is calculated by performing the complete-data ANOVA on the real Y data with initial guesses \tilde{y}_i filled in for y_i , $i = 1, \dots, m$; the residuals for the m missing values comprise the vector ρ .

These procedures work for the following reasons. The jk entry of B is

$$b_{jk} = \sum_{i=1}^n (z_{ij} - \hat{z}_{ij})(z_{ik} - \hat{z}_{ik}),$$

where z_{ij} and \hat{z}_{ij} (z_{ik} and \hat{z}_{ik}) are the observed and fitted values for unit i from the ANOVA of the j th (k th) missing value covariate on X . Now, $\sum_{i=1}^n x_{il}(z_{ik} - \hat{z}_{ik}) = 0$ for all X variables in the X matrix, by elementary properties of least squares estimates. Therefore, $\sum_{i=1}^n \hat{z}_{ij}(z_{ik} - \hat{z}_{ik}) = 0$ because \hat{z}_{ij} is a fixed linear combination of X variables $\{x_{il} : l = 1, \dots, p\}$ for unit i . Consequently,

$$b_{jk} = \sum_{i=1}^n z_{ij}(z_{ik} - \hat{z}_{ik}) = z_{jk} - \hat{z}_{jk},$$

the residual for the j th missing data covariate for the k th missing value because $z_{ij} = 1$ when $i = j$ and 0 otherwise. Similarly, the j th component of ρ is the sum over all n units of the residual for Y (with initial values filled in) times the residual for the j th missing-value covariate. By an argument completely analogous to that just given, this is simply the residual for the j th missing value.

Example 2.1 *Estimating Missing Values in a Randomized Block.* The following example of a randomized block design is taken from Cochran and Cox (1957, p. 111) and Rubin (1972, 1976b). Suppose that the two units, u_1 and u_2 , are missing as presented in Table 2.1. We formulate model (2.1) using a seven-dimensional parameter β consisting of five parameters for the means of the five treatments and two parameters for the block effects; the residual mean square is formed from the treatment by block interaction, with $(5 - 1) \times (3 - 1) = 8$ degrees of freedom when no data are missing.

Inserting the grand mean $\bar{y} = 7.7292$ for both missing values, we find the residual in the u_1 cell to be -0.0798 and in the u_2 cell to be -0.1105 . Thus, $\rho = -(0.0798, 0.1105)^T$. Also, we obtain the correct total sum of squares, $TSS_* = 1.1679$.

Inserting one for u_1 and zeros everywhere else, we find that the residual in the u_1 cell is 0.5333 and the residual in the u_2 cell is 0.0667. Similarly, inserting

Table 2.1 Example 2.1: strength index of cotton in a randomized block experiment

Treatments (pounds of potassium oxide per acre)	Blocks			Totals
	1	2	3	
36	u_1	8.00	7.93	15.93
54	8.14	8.15	7.87	24.16
72	7.76	u_2	7.74	15.50
108	7.17	7.57	7.80	22.54
144	7.46	7.68	7.21	22.35
Total	30.53	31.40	38.55	100.48

one for u_2 and zeros everywhere else, we find that the residual in the u_1 cell is 0.0667, and the residual in the u_2 cell is 0.5333. Hence

$$B = \begin{bmatrix} 0.5333 & 0.0667 \\ 0.0667 & 0.5333 \end{bmatrix} \quad \text{and} \quad B^{-1} = \begin{bmatrix} 1.9408 & -0.2381 \\ -0.2381 & 1.9408 \end{bmatrix}.$$

The least squares estimates of the missing values are

$$(\bar{y}, \bar{y}) - B^{-1}\rho = (7.8549, 7.9206)^T.$$

Thus, the least squares estimate of the u_1 cell is 7.8549 and that of the u_2 cell is 7.9206. The least squares estimates for the missing cells given by Cochran and Cox were found by an iterative method and agree with the values found here.

Estimated parameters based on the analysis of the filled-in data will be the correct least squares values. For example, the correct estimates of the treatment means are simply the treatment averages of observed and filled-in data (7.9283, 8.0533, 7.8069, 7.5133, 7.4500). Furthermore, the correct residual sum of squares, and thus the correct residual mean square s_*^2 , is obtained when the number of missing values m is subtracted from the residual degrees of freedom, $n - p$. However, other sums of squares generally will be too large, and standard errors generally will be too small.

2.7 Correct Least Squares Estimates of Standard Errors and One Degree of Freedom Sums of Squares

A simple extension of the technique of Section 2.6 yields the correct estimates of standard errors and one degree of freedom sums of squares. Let $\lambda = C^T\beta$,

where C is a vector of p constants, be a linear combination of β with estimate $\hat{\lambda} = C^T \hat{\beta}$ from the ANOVA of the data filled-in by least squares estimates. Because least squares estimates of the missing values have been filled in, $\hat{\beta} = \hat{\beta}_*$ and so $\hat{\lambda} = \hat{\lambda}_*$, the correct least squares estimate of λ . The standard error of $\hat{\lambda}$ obtained from the complete data ANOVA is

$$\text{SE} = s \sqrt{C^T (X^T X)^{-1} C}, \quad (2.17)$$

and the sum of squares attributable to λ from this analysis is

$$\text{SS} = \hat{\lambda}^2 / C^T (X^T X)^{-1} C. \quad (2.18)$$

The correct standard error of $\hat{\lambda} = \hat{\lambda}_*$ is, from Section 2.5.4,

$$\text{SE}_* = s_* \sqrt{C^T U C}, \quad (2.19)$$

and the correct sum of squares attributable to λ is

$$\text{SS}_* = \hat{\lambda}_*^2 / C^T U C. \quad (2.20)$$

Let H be the $(m \times 1)$ vector of complete-data ANOVA estimates of λ taking each of the m missing data covariates as the dependent variable rather than Y ; that is, in matrix terms

$$H^T = C^T (X^T X)^{-1} X^T Z. \quad (2.21)$$

Conveniently, H can be calculated at the same time B is being calculated: The i th component in H and the i th row in B are obtained from the complete-data ANOVA of the i th missing data covariate. Standard ANCOVA theory, or matrix algebra using results in Section 2.5.4, shows that

$$C^T U C = C^T (X^T X)^{-1} C + H^T B^{-1} H. \quad (2.22)$$

Equations (2.17), (2.19), (2.21), (2.22), and the fact that

$$s_*^2 = s^2(n - p)/(r - p)$$

imply that SE_* can be simply expressed in terms of output from the complete-data ANOVA:

$$\text{SE}_* = \sqrt{\frac{n-p}{r-p} (\text{SE}^2 + s^2 H^T B^{-1} H)}. \quad (2.23)$$

Similarly, (2.18), (2.20) with $\lambda = \lambda_*$, (2.21), and (2.22) imply that SS_* can be simply expressed in terms of output from the complete-data ANOVA:

$$\text{SS}_* = \text{SS}/(1 + (\text{SS}/\hat{\lambda}^2) H^T B^{-1} H). \quad (2.24)$$

Example 2.2 Adjusting Standard Errors for Filled-In Missing Values (Example 2.1 Continued). To apply the method just described, $m + 2$ complete data ANOVA's are required: an initial one on starting filled-in Y data, one for each of the m missing data covariates, and a final ANOVA on the least squares filled-in Y data. Following Rubin (1976b), we consider the data in Table 2.1 and the linear combination of parameters that correspond to contrasting treatment 1 and treatment 2; in terms of the parameterization of Example 2.1, $C^T = (1, -1, 0, 0, 0, 0, 0)$ and $X^T X$ is a 7×7 block diagonal matrix whose upper left 5×5 submatrix is diagonal with all elements equal to 3. Thus, $\hat{\lambda}$ is simply the mean Y -values of the three units having treatment 1 minus the mean Y -values of the three units having treatment 2, with associated complete data standard error $s\sqrt{2/3}$ and sum of squares $3\hat{\lambda}^2/2$.

As in Example 2.1, for the initial ANOVA, estimate both missing values by the grand mean to obtain residuals $\rho = (-0.0798, -0.1105)$ and the correct total sum of squares, $TSS_* = 1.1679$. For $i = 1, 2, \dots, m$, fill in 1 for the i th missing value and set all other values to zero and analyze the resultant missing data covariate by the complete-data ANOVA program: r_i is the vector of residuals corresponding to the m missing values, and h_i is the estimate of the linear combination of parameters being tested. The resultant B for our example is given in Example 2.1, the resultant H^T is $(0.3333, 0.0000)$, and consequently, the resultant $H^T B^{-1} H$ is 0.2116.

Now fill in the least squares estimates $(7.8549, 7.9206)$ of the missing values, as found in Example 2.1, and compute the ANOVA on the filled-in data. The resultant estimate of λ is $\hat{\lambda} = -0.1250$, with $s^2 = 0.0368$, $SE = 0.1567$, and $SS = 0.0235$. From (2.23), the correct estimated standard error of $\hat{\lambda}$ is

$$SE_* = \sqrt{(8/6)(0.0246 + 0.0368 \times 0.2116)} = 0.2077,$$

and from (2.24), the correct sum of squares attributable to λ is

$$SS_* = 0.0235/(1 + 1.5 \times 0.2116) = 0.0178.$$

2.8 Correct Least-Squares Sums of Squares with More Than One Degree of Freedom

A generalization of the technique of Section 2.7 yields the correct sum of squares with more than one degree of freedom. The technique presented here is due to Rubin (1976b), the related earlier work includes Tocher (1952) and Wilkinson (1958b); later work includes Jarrett (1978).

Let $\lambda = C^T \beta$, where C is a $p \times w$ matrix of constants, be w linear combinations of β for which the sum of squares is desired, and let $\hat{\lambda}_* = C^T \hat{\beta}_*$ be the correct least squares estimate of λ . When least squares estimates of the missing values

have been filled in, $\hat{\beta} = \hat{\beta}_*$, and thus, $\hat{\lambda} = \hat{\lambda}_*$. We suppose for simplicity that the w linear combinations have been chosen to be orthonormal with complete data, in the sense that

$$C^T(X^T X)^{-1} C = I_w. \quad (2.25)$$

That is, with complete data, the covariance matrix of $\hat{\lambda}$ is $\sigma^2 I_w$. Thus, the sum of squares attributable to λ from the complete-data ANOVA is

$$SS = \hat{\lambda}^T \hat{\lambda}. \quad (2.26)$$

The correct sum of squares to attribute to λ is

$$SS_* = \hat{\lambda}_*^T (C^T U C)^{-1} \hat{\lambda}_*. \quad (2.27)$$

Letting H be the $m \times w$ matrix of complete-data ANOVA estimates of λ for the m missing-data covariates, standard ANCOVA theory, or matrix algebra using results in Section 2.5.4, shows that (2.22) holds in general; hence, because the components of $\hat{\lambda}$ are orthonormal and $\hat{\lambda} = \hat{\lambda}_*$ with least squares estimates for missing values,

$$SS_* = \hat{\lambda}^T (I + H^T B^{-1} H)^{-1} \hat{\lambda}, \quad (2.28)$$

or using Woodbury's identity (Rao 1965, p. 29) and (2.26),

$$SS_* = SS - (H\hat{\lambda})^T (HH^T + B)^{-1} (H\hat{\lambda}). \quad (2.29)$$

Equation (2.28) involves the inversion of a $w \times w$ symmetric matrix, whereas (2.29) involves the inversion of an $m \times m$ matrix. Consequently, (2.28) is computationally preferable when $w < m$.

Example 2.3 *Adjusting Sums of Squares for the Filled-In Values (Example 2.2 Continued).* The treatment sum of squares has four degrees of freedom, which we span with the following orthonormal contrasts of five treatment means:

$$\sqrt{\frac{3}{20}}(4, -1, -1, -1, -1, 0, 0),$$

$$\sqrt{\frac{1}{4}}(0, 3, -1, -1, -1, 0, 0),$$

$$\sqrt{\frac{1}{2}}(0, 0, 2, -1, -1, 0, 0),$$

$$\sqrt{\frac{3}{2}}(0, 0, 0, 1, -1, 0, 0).$$

Table 2.2 Example 2.3: the corrected analysis of variance on the filled-in data

Source of variation	df	SS	MS	F
Blocks, unadjusted	2	0.0977		
Treatments, adjusted for blocks	4	0.7755	0.1939	3.9486
Error	6	0.2947	0.0491	
Total	12	1.1679		

Treatment means: (7.9283, 8.0533, 7.8069, 7.5133, 7.4500).

Contrast: treatment 1 – treatment 2 = –0.1250, SE = 0.2077.

Note that with complete data, the linear combinations have covariance matrix $\sigma^2 I$.

The values of the four contrasts obtained from the complete-data ANOVA of the first missing-data covariate gives the first row of H , and the complete-data ANOVA of the second missing data covariate gives the second row of H :

$$H = \begin{bmatrix} 0.5164 & 0.0000 & 0.0000 & 0.0000 \\ -0.1291 & -0.1667 & 0.4714 & 0.0000 \end{bmatrix}.$$

Thus, H is calculated at the same time B is calculated.

From the final complete-data ANOVA of the data with least squares estimates filled in, we have $SS = 0.8191$, and $\hat{\lambda}^T = (0.3446, 0.6949, 0.4600, 0.0775)$. From (2.29), $SS_* = 0.7755$.

A summary of the resulting ANOVA for this example appears in Table 2.2, where the blocks sum of squares (unadjusted for treatments) has been found by subtracting the corrected treatment and error sums of squares (0.7755 and 0.2947) from the corrected total sum of squares (1.1679) found in Example 2.1.

Problems

- 2.1 Review the literature on missing values in ANOVA from Allan and Wishart (1930) through Dodge (1985).
- 2.2 Prove that $\hat{\beta}$ in (2.2) is (a) the least squares estimate of β , (b) the minimum variance unbiased estimate, and (c) the maximum likelihood estimate under normality. Which of these properties does s^2 possess, and why?
- 2.3 Outline the distributional results leading to (2.6) being distributed as F .

- 2.4** Summarize the argument that Bartlett's ANCOVA method leads to correct least squares estimates of missing values.
- 2.5** Prove that (2.12) follows from the definition of U^{-1} .
- 2.6** Provide intermediate steps leading to (2.13), (2.14), and (2.15).
- 2.7** Using the notation and results of Section 2.5.4, justify (2.16) and the method for calculating B and ρ that follows it.
- 2.8** Carry out the computations leading to the results of Example 2.1.
- 2.9** Justify (2.17)–(2.20).
- 2.10** Show (2.22) and then (2.23) and (2.24).
- 2.11** Carry out the computations leading to the results of Example 2.2.
- 2.12** Carry out the computations leading to the results of Example 2.3.
- 2.13** Carry out a standard ANOVA for the following data, where three values have been deleted from a (5×5) Latin square (Snedecor and Cochran 1967, p. 313).

Row	Yields (grams) of plots of millet arranged in a latin square ^a				
	Column				
Row	1	2	3	4	5
1	B: –	E: 230	A: 279	C: 287	D: 202
2	D: 245	A: 283	E: 245	B: 280	C: 260
3	E: 182	B: –	C: 280	D: 246	A: 250
4	A: –	C: 204	D: 227	E: 193	B: 259
5	C: 231	D: 271	B: 266	A: 334	E: 338

^aSpacings (in.): A, 2; B, 4; C, 6; D, 8; and E, 10.

3

Complete-Case and Available-Case Analysis, Including Weighting Methods

3.1 Introduction

In Chapter 2, we discussed the analysis of data with missing values confined to a single outcome variable that is related to completely observed predictor variables through a linear model. We now discuss the more general problem with values missing for more than one variable. In this chapter, we discuss “complete-case” (CC) analysis, which confines the analysis to the set of units with no missing values and modifications and extensions. In the following two chapters, we discuss imputation methods. Afifi and Elashoff (1966) review the earlier literature on missing data, including some of the methods discussed here. Although these methods appear in statistical computing software and are still widely used, we do not generally recommend any of them except in situations where the amount of additional missing information in the incomplete units is limited. The procedures in Part II provide sounder solutions in more general circumstances.

3.2 Complete-Case Analysis

CC analysis confines attention to cases (units) where all the variables are present. Advantages of this approach are (i) simplicity because standard complete-data statistical analyses can be applied without modifications and (ii) comparability of univariate statistics because these are all calculated on a common sample base of units. If the additional information for the target parameters in the incomplete units is small, then including them in the analysis yields limited gains; as discussed in the following, the amount of this additional information depends on the fraction of incomplete cases, which values are missing, the missingness mechanism, and the specifics of the analysis. Disadvantages stem from the potential loss of information in discarding incomplete

units. This loss of information has two aspects: loss of precision and bias when the missingness mechanism is not missing completely at random (MCAR), so that the complete units are not a random sample of all the units. The degree of bias and loss of precision depends not only on the fraction of complete units and pattern of missing data but also on the extent to which complete and incomplete units differ and on the estimands of interest.

We focus first on precision rather than on bias. Let $\hat{\theta}_{CC}$ be an estimate of a scalar parameter θ from the complete units, and let $\hat{\theta}_{EFF}$ be an efficient estimate of θ based on all the available data, for example the maximum likelihood (ML) estimate under a particular model. A measure of the loss of efficiency of $\hat{\theta}_{CC}$ is then Δ_{CC} , where

$$\text{Var}(\hat{\theta}_{CC}) = \text{Var}(\hat{\theta}_{EFF})(1 + \Delta_{CC}). \quad (3.1)$$

Example 3.1 *Efficiency of Complete-Case Analysis for Bivariate Normal Monotone Data.* Consider bivariate normal monotone data, where r of the n cases are complete and $n - r$ have Y_1 observed but Y_2 missing. Suppose that the mean of Y_j is estimated by the CC mean \bar{y}_j^{CC} , and the missing data are MCAR, so the bias of CC analysis is not an issue. For estimating the mean of Y_1 , dropping the incomplete units on Y_1 translates directly to a loss in sample size:

$$\Delta_{CC}(\bar{y}_1^{CC}) = \frac{n - r}{r},$$

so if half the units are missing Y_2 , the sampling variance is doubled. For the mean of Y_2 , the loss of efficiency of CC analysis depends not only on the fraction of missing units but also on the squared correlation, ρ^2 , between Y_1 and Y_2 :

$$\Delta_{CC}(\bar{y}_2^{CC}) \approx \frac{(n - r)\rho^2}{n(1 - \rho^2) + r\rho^2}. \quad (3.2)$$

(The derivation of this expression is in Section 7.2.) Thus, $\Delta_{CC}(\bar{y}_2^{CC})$ ranges from zero, when Y_1 and Y_2 are uncorrelated, to $(n - r)/r$, as ρ^2 tends to one. For the coefficients of the regression of Y_2 on Y_1 , CC analysis is fully efficient (that is $\Delta_{CC} = 0$) because the $n - r$ incomplete observations on Y_1 provide no information for these parameters.

The potential bias of CC analysis also depends on the nature of the analysis.

Example 3.2 *Bias of Complete-Case Inferences for Means.* For inference about a mean, the bias depends on the fraction of incomplete units and the extent to which complete and incomplete units differ on the variable of interest. Specifically, suppose a variable Y has missing values, and partition the population into strata consisting of respondents and nonrespondents to Y , with corresponding

proportions π_{CC} and $1 - \pi_{CC}$, respectively. Let μ_{CC} and μ_{IC} denote the population means of Y in these strata, that is, of the complete and incomplete units, respectively. The overall mean can be written as $\mu = \pi_{CC}\mu_{CC} + (1 - \pi_{CC})\mu_{IC}$, and hence, the bias of the CC sample mean is

$$\mu_{CC} - \mu = (1 - \pi_{CC})(\mu_{CC} - \mu_{IC}),$$

the expected fraction of incomplete units multiplied by the differences in the means for complete and incomplete units. Under MCAR, $\mu_{CC} = \mu_{IC}$ and the bias is zero.

Example 3.3 Bias and Precision of Complete-Case Inferences for Regression Coefficients. Consider estimation of the regression of Y on X_1, \dots, X_p from data with missing values on Y and/or the X 's, where the regression function is correctly specified. The CC estimates of the regression coefficients are not subject to bias if the probability of being a complete unit depends on X_1, \dots, X_p but not on Y , because the analysis conditions on the values of the covariates (Glynn and Laird 1986). This class of mechanisms includes missing not at random (MNAR) mechanisms where the probability that a covariate is missing depends on the values of missing covariates. The CC estimates of the regression coefficients are biased if the probability of being complete depends on Y after conditioning on the covariates.

Concerning precision, suppose for simplicity that Y and X_2, \dots, X_p are fully observed, and missing values are confined to X_1 . The incomplete units may provide little information for the estimate of the regression coefficient of X_1 because this involves the partial association between Y and X_1 given X_2, \dots, X_p , and only one of the values in the pair (Y, X_1) is observed. On the other hand, the incomplete units provide considerable information for the intercept and regression coefficients of X_2, \dots, X_p on Y because both variables involved in these coefficients are observed (Little 1992). Methods for recovering this information are discussed in Part II of this book. The implication is that if the coefficient of primary interest is for a variable that is missing for the incomplete units, the loss of efficiency from CC analysis may be minor.

Example 3.4 Bias and Precision of Complete Case Inferences for an Odds Ratio. Even milder restrictions on the relationship between missingness and the measured variables apply to certain other analyses. For example, if Y_1 and Y_2 are dichotomous and inference concerns the odds ratio in the 2×2 table of counts classified by Y_1 and Y_2 , then CC analysis is not subject to possible bias if the logarithm of the probability of response is an additive function of Y_1 and Y_2 (Kleinbaum et al. 1981). This result underpins the possible validity of “case-control” studies for estimating odds ratios from nonrandomized observational

studies. In terms of precision, supplemental margins on Y_1 and Y_2 provide little information for the odds ratio but can reduce bias and increase precision for estimating the marginal distributions of these variables, which can be of substantial interest.

The discarded information from incomplete units can be used to study whether the complete units are plausibly a random subsample of the original sample, that is, whether MCAR is a reasonable assumption. A simple procedure is to compare the distribution of a particular variable Y_j based on complete units with the distribution of Y_j based on incomplete units for which Y_j is recorded. Significant differences indicate that the MCAR assumption is invalid, and CC analysis yields potentially biased estimates. Such tests are useful but have limited power when the sample of incomplete units is small. Also, the tests can offer no direct evidence on the validity of the weaker missing at random (MAR) assumption.

A strategy for adjusting for the bias in CC analysis is to assign unit weights for use in subsequent analyses. This strategy is common for unit nonresponse in sample surveys, where all the survey items are missing for units in the sample that did not participate. Information available for respondents and nonrespondents, such as their geographic location, can be used to assign weights to the respondents that, at least partially, adjust for nonresponse bias.

3.3 Weighted Complete-Case Analysis

3.3.1 Weighting Adjustments

In this section, we consider a modification of CC analysis that differentially weights the complete units to adjust for bias. The basic idea is closely related to weighting in randomization inference for finite population surveys. The next example reviews the basic elements of that approach to inference.

Example 3.5 *Randomization Inference in Surveys with Complete Response.* Suppose inferences are required for characteristics of a population with N units, and let $Y = (y_{ij})$, where $y_i = (y_{i1}, \dots, y_{iK})$ represents a vector of K items for unit i , $i = 1, \dots, N$. For unit i , define the sample indicator function

$$I_i = \begin{cases} 1, & \text{unit } i \text{ included in the sample,} \\ 0, & \text{unit } i \text{ not included in the sample,} \end{cases}$$

and let $I = (I_1, \dots, I_N)^T$. Sample selection processes can be characterized by a distribution for I given Y and design information Z . Randomization inference

generally requires that units be selected by *probability sampling*, which is characterized by the following two properties:

1. The sampler determines the distribution before any Y -values are known. In particular, $f(I \mid Y, Z) = f(I \mid Z)$, because the distribution cannot depend on the unknown values of items Y to be sampled in the survey. Such a mechanism is called “unconfounded” (Rubin 1987a, chapter 2).
2. Every unit has a positive (known) probability of selection. Writing $\pi_i = E(I_i \mid Y, Z) = \Pr(I_i = 1 \mid Y, Z)$, we require that $\pi_i > 0$ for all i . In *equal probability* sample designs, such as simple random sampling, this probability is the same for all units.

For example, if Z is a variable defining strata, recorded for all units in the population, then *stratified random sampling* takes a simple random sample of n_j units from the N_j population units in stratum $Z = j, j = 1, \dots, J$. The distribution of the sampling indicator is then

$$f(I \mid Y, Z) = f(I \mid Z) = \begin{cases} \prod_{j=1}^J \binom{N_j}{n_j}^{-1}, & \text{if } \sum_{i:z_i=j} I_i = n_j \text{ for all } j, \\ 0, & \text{otherwise,} \end{cases}$$

where $\binom{N_j}{n_j}$ is the number of ways n_j units can be chosen from stratum j .

Detailed treatments of randomization inference for surveys can be found in sampling theory texts such as Cochran (1977), Hansen et al. (1953), or Lohr (2010). In outline, let Y_{inc} denote the set of Y -values included in the sample (that is, with $I_i = 1$), and let T denote a population quantity of interest. The following steps are involved in deriving inferences for T :

- (a) The choice of a statistic $t(Y_{\text{inc}})$, a function of the sampled values Y_{inc} , that is (approximately) unbiased for T in repeated samples. For example, if $T = \bar{Y}$, the population mean, then an unbiased estimate of T from a stratified random sample is the stratified sample mean

$$t = \bar{y}_{\text{st}} = \frac{1}{N} \sum_{j=1}^J N_j \bar{y}_j,$$

where \bar{y}_j is the sample mean in stratum j .

- (b) The choice of a statistic $v(Y_{\text{inc}})$ that is (approximately) unbiased for the sampling variance of $t(Y_{\text{inc}})$ in repeated sampling, that is, treating all population values as fixed and the sampling indicator I as random. For example, under

stratified random sampling, it can be shown that the sampling variance of \bar{y}_{st} is

$$\text{Var}(\bar{y}_{st}) = \frac{1}{N^2} \sum_{j=1}^J N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j} \right) S_j^2,$$

where S_j^2 is the population variance of the values of Y in stratum j (with denominator N_j). The statistic

$$v(Y_{inc}) = \frac{1}{N^2} \sum_{j=1}^J N_j^2 \left(\frac{1}{n_j} - \frac{1}{N_j} \right) s_j^2,$$

where s_j^2 is the variance of the sampled values of Y in stratum j , is an unbiased estimate of $\text{Var}(\bar{y}_{st})$ under stratified random sampling.

- (c) The calculation of interval estimates for T , assuming t has an approximate normal sampling distribution over all stratified random samples. For example, a large sample 95% confidence interval for \bar{Y} under stratified random sampling is given by $C_{95}(\bar{Y}) = \bar{y}_{st} \pm 1.96\sqrt{v(Y_{inc})}$, where 1.96 is the 97.5 percentile of the normal distribution. The normal approximation is justified by appealing to a finite population version of the central limit theorem (Hajek 1960 or Li and Ding 2017).

Because the population values of Y are treated as fixed, an attractive aspect of the randomization approach is the avoidance of a model specification for the population Y -values, although the confidence interval in (c) requires that the distribution of Y -values in the population is sufficiently well behaved for the sampling distribution of t to be approximately normal. An alternative approach to inference about finite population quantities is to specify, in addition to the distribution of I , a model for Y , often in the form of a density $f(Y|Z, \theta)$ indexed by unknown vector parameter θ . This model is then used to predict the non-sampled values of Y , and thereby population characteristics that are functions of sampled and nonsampled values. This model-based approach is applied to incomplete data problems in Parts II and III of the book.

One way of viewing probability sampling is that a unit selected from a target population with probability π_i is “representing” π_i^{-1} units in the population and hence should be given weight π_i^{-1} when estimating population quantities. For example, in a stratified random sample, a selected unit in stratum j represents N_j/n_j population units. The population total T can be estimated by the weighted sum

$$t_{HT} = \sum_{i=1}^n y_i \pi_i^{-1},$$

which is called the Horvitz–Thompson estimate (Horvitz and Thompson 1952). The stratified mean can be written in the form

$$\bar{y}_{st} \equiv \bar{y}_w = \frac{1}{n} \sum_{i=1}^n w_i y_i, \quad (3.3)$$

where $w_i = n\pi_i^{-1} / \sum_{k=1}^n \pi_k^{-1}$ is the sampling weight attached to the i th unit, scaled to sum to the sample size n . The estimate \bar{y}_w is unbiased for the mean of Y in stratified random samples and is approximately unbiased in other designs.

Of course, t_{HT} and \bar{y}_w can only be calculated with complete response. Weighting class estimators extend this approach when there is nonresponse. If the probabilities of response for each responding unit i , say ϕ_i , were known, then

$$\Pr(\text{selection and response}) = \Pr(\text{selection}) \times \Pr(\text{response} \mid \text{selection}) = \pi_i \phi_i$$

and (3.3) can be replaced by

$$\bar{y}_w = \frac{1}{r} \sum_{i=1}^r w_i y_i, \quad (3.4)$$

where the sum is now over responding units i and $w_i = (r/(\pi_i \phi_i)) / \sum_{k=1}^r (\pi_k \phi_k)^{-1}$. In practice, the response probability ϕ_i is not known and needs to be estimated based on information available for respondents and nonrespondents. The simplest approach is illustrated by Example 3.6, and a more general approach is illustrated by Example 3.7.

Example 3.6 Weighting Class Estimator of the Mean. Suppose we partition the sample into J “weighting classes” on the basis of variables observed for respondents and nonrespondents. Let C denote this weighting class variable. If n_j is the sample size, r_j is the number of respondents in weighting class $C = j$, and $r = \sum_{j=1}^J r_j$, a simple estimator of the response probability for units in class j is r_j/n_j . Then responding units in weighting class j receive weight

$$w_i = r(\pi_i \hat{\phi}_i)^{-1} / \sum_{k=1}^r (\pi_k \hat{\phi}_k)^{-1}, \quad \text{where } \hat{\phi}_i = r_j/n_j \text{ for units } i \text{ in class } j. \quad (3.5)$$

If the sampling weight is not constant within a weighting class, some authors advocate including sampling weights in the estimate of the response probability, but this approach does not generally correct for bias when the sampling weight is related to both C and the outcome (Little and Vartivarian 2003). As discussed in that article, a better approach is to incorporate design information in the

formation of weighting classes. The weighting class estimate of the mean is then given by (3.4) with weights given by (3.5). For equal probability designs, π_i is a constant, and this estimate can be written in the simpler form:

$$\bar{y}_{wc} = n^{-1} \sum_{j=1}^J n_j \bar{y}_{jR}, \quad (3.6)$$

where \bar{y}_{jR} is the respondent mean in class j , and $n = \sum_{j=1}^J n_j$ is the total sample size.

This estimate is unbiased under the following form of the MAR assumption, which Oh and Scheuren (1983) call *quasirandomization*, by analogy with random sampling for selection of units:

Assumption 3.1 Quasirandomization. Respondents in weighting class j are a random sample of the sampled units (that is the data are MCAR within adjustment class j).

Weighting classes can be formed from survey design variables or from sampled items recorded for both respondents and nonrespondents. Weighting class adjustments are used primarily to handle unit nonresponse, where none of the sampled items are recorded for nonrespondents. In these applications, only survey design variables are available for forming adjustment classes. Adjustment classes should ideally be chosen so that (i) Assumption 3.1 is satisfied and (ii) under Assumption 3.1, the mean squared error of estimates such as \bar{y}_{wc} is minimized.

Weighting class adjustments are simple because the same weights are obtained regardless of the survey outcome Y to which they are applied. Thus in large surveys, with MAR and hundreds of Y s, bias is handled by a single set of weights. On the other hand, this simplicity entails a cost, in that weighting is inefficient and generally involves an increase in sampling variance for outcomes that are weakly related to the weighting class variable. A simple formula for the increase in sampling variance of a sample mean in such situations is obtained by assuming random sampling within weighting classes, ignoring sampling variation in the weights, and assuming that an outcome Y has constant variance σ^2 . Then, if the weights are scaled to average 1:

$$\text{Var}\left(\frac{1}{r} \sum_{i=1}^r w_i y_i\right) \doteq \frac{\sigma^2}{r^2} \left(\sum_{i=1}^r w_i^2 \right) = \frac{\sigma^2}{r} (1 + \text{cv}^2(w_i)), \quad (3.7)$$

where $\text{cv}(w_i)$ is the coefficient of variation of the weights. Thus, the squared coefficient of variation of the weights is a rough measure of the proportional increase in sampling variance due to weighting (Kish 1992).

In situations where the weighting class variable is predictive of Y , Eq. (3.7) no longer applies, and weighting can in fact lead to a reduction in sampling variance. Oh and Scheuren (1983) derive the sampling variance of (3.6) under simple random sampling and propose the following estimate of the mean squared error of \bar{y}_{wc} :

$$\widehat{\text{mse}}(\bar{y}_{wc}) = \sum_{j=1}^J \left(\frac{n_j}{n} \right)^2 \left(1 - \frac{r_j n}{n_j N} \right) \frac{s_{jR}^2}{r_j} + \frac{N-n}{(N-1)n^2} \sum_{j=1}^J n_j (\bar{y}_{jR} - \bar{y}_{wc})^2, \quad (3.8)$$

where s_{jR}^2 is the variance of sampled and responding units in class j ; $100(1-\alpha)\%$ confidence intervals for the population mean may be constructed of the form $\bar{y}_{wc} \pm z_{1-\alpha/2} \{ \widehat{\text{mse}}(\bar{y}_{wc}) \}^{1/2}$, where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution.

Little and Vartivarian (2005) also discuss the change in mean squared error from weighting. A simple qualitative summary of the results is shown in Table 3.1, which indicates the effect of weighting on the bias and sampling variance of an estimated mean, according to whether the associations between the adjustment cells and the outcome and missing indicator are high or low. When the adjustment cell variable is weakly associated with the survey outcome (cells LL and HL), weighting has little effect on bias; it actually makes things worse when the adjustment cell variable is related to nonresponse, by increasing sampling variance (cell HL), a reflection of variable weights. When the adjustment cell variable is strongly associated with the outcome (cells LH and HH), weighting tends to reduce sampling variance and also reduces bias when the adjustment cell variable is strongly related to nonresponse (cell HH). Clearly, weighting is only effective for outcomes that are associated with the adjustment cell variable because otherwise it increases the sampling variance with no compensating reduction in bias.

Table 3.1 Example 3.6: effect of weighting adjustments on bias and sampling variance of a mean, by strength of association of the adjustment cell variables with nonresponse and outcome

		Association with outcome	
Association with nonresponse		Low (L)	High (H)
Low (L)	Bias: —	Bias: —	Bias: —
	Var: —	Var: ↓	Var: ↓
High (H)	Bias: —	Bias: —	Bias: ↓
	Var: ↑	Var: ↓	Var: ↓

In summary, bias is reduced by choosing adjustment classes that are predictive of both response and Y . Sampling variance is reduced by choosing adjustment classes that are predictive of Y so that the within-class variance of Y is reduced. Adjustment cells should be formed that avoid small respondent sample sizes, which lead to large weights.

Example 3.7 Propensity Weighting. Let X denote the set of variables observed for both respondents and nonrespondents. Weighting class estimates can be applied in practice when the set of variables X is limited. However, in some settings, such as panel surveys when information from a prior survey is available for nonrespondents, joint classification by all the recorded variables is not practical because the number of weighting classes becomes too large, and includes cells with nonrespondents but no respondents, for which the nonresponse weight is infinite. The theory of propensity scores (Rosenbaum and Rubin 1983, 1985), discussed in the context of survey nonresponse in Rubin (1985a), Little (1986), and Czajka et al. (1992), provides a prescription for choosing the coarsest reduction of X to a weighting class variable C so that Assumption 3.1 is approximately satisfied. Suppose that the data are MAR, that is

$$\Pr(M | X, Y, \phi) = \Pr(M | X, \phi), \quad (3.9)$$

where ϕ are unknown parameters, so that Assumption 3.1 is satisfied when C is chosen to be X . Define the nonresponse propensity for unit i

$$\rho(x_i, \phi) = \Pr(m_i = 1 | \phi),$$

and assume that this is strictly positive for all values of x_i . Then

$$\begin{aligned} \Pr(m_i = 0 | y_i, \rho(x_i, \phi), \phi) &= E(\Pr(m_i = 0 | y_i, x_i, \phi) | y_i, \rho(x_i, \phi), \phi) \\ &= E(\Pr(m_i = 0 | x_i, \phi) | y_i, \rho(x_i, \phi), \phi), \text{ by Eq. (3.9)} \\ &= E(\rho(x_i, \phi) | y_i, \rho(x_i, \phi), \phi), \text{ by definition of } \rho(x_i, \phi) \\ &= \rho(x_i, \phi), \text{ for all } x_i. \end{aligned}$$

Hence,

$$\Pr(M | \rho(X, \phi), Y, \phi) = \Pr(M | \rho(X, \phi), \phi),$$

so that respondents are a random subsample within strata defined by the propensity score $\rho(X, \phi)$.

In practice, the parameter ϕ in $\rho(X, \phi)$ is unknown and needs to be estimated from sample data. A practical procedure is to (i) estimate $\rho(X, \phi)$ as $\rho(X, \hat{\phi})$, where $\hat{\phi}$ is estimated from a logistic, probit, or robit (Liu 2005) regression of the missing-data indicator M on X , based on respondent and nonrespondent data; (ii) form a grouped variable by coarsening $\rho(X, \hat{\phi})$ into

five to ten values; and (iii) let C equal that grouped variable, so that within adjustment class j , all respondents and nonrespondents have the same value of the grouped propensity score. A variant of this procedure is to weight respondents i directly by the inverse of the estimated propensity score $\rho(X, \hat{\phi})^{-1}$ (Cassel et al. 1983). Note that weighting class estimation is a special case of this method, where X is a single categorical variable, and the logistic model of M on X is saturated. Under the modeling assumptions underlying $\Pr(M | X, \phi)$, this method removes nonresponse bias, but it may yield estimates with extremely high sampling variance because respondents with very low estimated response propensities receive large nonresponse weights and may be unduly influential when estimating means and totals. Also, weighting directly by $\rho(X, \hat{\phi})^{-1}$ places more reliance on correct model specification of the regression of M on X than response propensity stratification, which uses $\rho(X, \hat{\phi})$ to form adjustment classes.

Example 3.8 *Inverse Probability Weighted Generalized Estimating Equations.* More generally, let $y_i = (y_{i1}, \dots, y_{iK})$ denote a vector of variables for unit i subject to missing values, and suppose y_i is fully observed for $i = 1, \dots, r$ and y_i is missing or partially observed for $i = r+1, \dots, n$. Define $m_i = 1$ if y_i is incomplete and $m_i = 0$ if y_i is complete. Let $x_i = (x_{i1}, \dots, x_{ip})^T$ denote a vector of fully observed covariates, and suppose that interest concerns the mean of the distribution of y_i given x_i , which is assumed to have the form $g(x_i, \beta)$, where g is a (possibly nonlinear) regression function indexed by an unknown parameter β of dimension d . Furthermore, let $z_i = (z_{i1}, \dots, z_{iq})^T$ be a vector of fully observed auxiliary variables that potentially predict whether or not y_i is complete but are not included in the regression model for y_i given x_i . If there were no missing values, the solution of the generalized estimating equation (GEE)

$$\sum_{i=1}^n D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0, \quad (3.10)$$

where $D_i(x_i, \beta)$ is a suitably chosen $(d \times K)$ matrix of known functions of x_i provides a consistent estimate of β under mild regularity conditions (Liang and Zeger 1986). With missing data, CC analysis replaces Eq. (3.10) by

$$\sum_{i=1}^r D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0, \quad (3.11)$$

which yields consistent estimates provided that

$$\Pr(m_i = 1 | x_i, y_i, z_i, \phi) = \Pr(m_i = 1 | x_i, \phi), \quad (3.12)$$

so that missingness does not depend on y_i or z_i after conditioning on x_i . Inverse-probability weighted generalized estimating equation (IPWGEE, see Robins et al. 1995) replaces Eq. (3.11) by

$$\sum_{i=1}^r w_i(\hat{\alpha}) D_i(x_i, \beta)(y_i - g(x_i, \beta)) = 0, \quad (3.13)$$

where $w_i(\hat{\alpha}) = 1/p(x_i, z_i | \hat{\alpha})$, with $p(x_i, z_i | \hat{\alpha})$ an estimate of the probability of being a complete unit, obtained for example by a logistic regression of m_i on x_i and z_i . Here, α is the vector parameter of the logistic regression, estimated, for example by ML. If this regression is correctly specified, Eq. (3.13) yields a consistent estimate of β provided that

$$\Pr(m_i = 1 | x_i, y_i, z_i, \phi) = \Pr(m_i = 1 | x_i, z_i, \phi),$$

which is less restrictive than Eq. (3.12) because missingness is allowed to depend on z_i as well as x_i . Thus, IPWGEE can correct for the bias of unweighted GEE attributable to dependence of the missingness mechanism on z_i . Robins et al. (1995) discuss sampling variance estimation for IPWGEE estimates and extensions to monotone and nonmonotone patterns of missingness. Additional references to this approach include Manski and Lerman (1977), Zhao and Lipsitz (1992), Park (1993), Robins and Rotnitzky (1995), and Lipsitz et al. (1999).

Augmented IPWGEE is an extension of IPWGEE that creates predictions from a model to recover information in the incomplete units and applies IPWGEE to the residuals from the model. There are also extensions of IPWGEE to missing not at random models. See in particular the paper by Scharfstein et al. (1999) and the discussion of that paper. As discussed in Chapter 15, there is never any direct empirical evidence against MAR without assumptions; therefore, it is wise to consider such assumptions carefully.

3.3.2 Poststratification and Raking to Known Margins

In the weighting class estimate (3.6), the proportion N_j/N of the population in weighting class j is estimated by the sample proportion, n_j/n . For the methods of this section, we assume that information about the weighting class proportions is available from external sources such as a larger survey or census.

Example 3.9 Poststratification. Suppose the population proportions N_j/N are known from external sources. In that case, an alternative to the weighting class estimate is the post stratified mean

$$\bar{y}_{ps} = \frac{1}{N} \sum_{j=1}^J N_j \bar{y}_{jR}. \quad (3.14)$$

Under Assumption 3.1, \bar{y}_{ps} is unbiased for \bar{Y} with sampling variance

$$\text{Var}(\bar{y}_{\text{ps}}) = \frac{1}{N^2} \sum_{j=1}^J N_j^2 \left(1 - \frac{r_j}{N_j}\right) \frac{S_{jR}^2}{r_j}. \quad (3.15)$$

An estimate of (3.15) is obtained by replacing the population variance S_{jR}^2 in Eq. (3.15) with the sample variance for respondents in class j , s_{jR}^2 . In most circumstances, \bar{y}_{ps} has lower mean squared error than \bar{y}_{wc} , except when the respondent sample sizes r_j and the between-class variance of Y are small (Holt and Smith 1979). For further discussions of poststratification and extensions, see Little (1993b), Lazzeroni and Little (1998), Bethlehem (2002), and Gelman and Carlin (2002).

Example 3.10 Raking Ratio Estimation. Suppose that the weighting classes are defined by the joint levels of two cross-classifying factors X_1 and X_2 , with J and L levels, respectively, and suppose that $n_{j\ell}$ units of N_{jl} in the population are sampled in the class with $X_1 = j$, $X_2 = \ell$, for $j = 1, \dots, J$, $\ell = 1, \dots, L$. The value of a variable Y is observed for $r_{j\ell}$ out of the $n_{j\ell}$ sampled units in class (j, ℓ) . The poststratified and weighting class estimates take the form

$$\bar{y}_{\text{ps}} = \frac{1}{N} \sum_{j=1}^J \sum_{\ell=1}^L N_{j\ell} \bar{y}_{j\ell R},$$

$$\bar{y}_{\text{wc}} = \frac{1}{n} \sum_{j=1}^J \sum_{\ell=1}^L n_{j\ell} \bar{y}_{j\ell R},$$

respectively, where $\bar{y}_{j\ell R}$ is the mean of responding units in class $j\ell$, with $X_1 = j$ and $X_2 = \ell$. An intermediate estimate can be based on the respondent cell means when the marginal counts for X_1 and X_2 , namely $N_{j+} = \sum_{\ell=1}^L N_{j\ell}$ and $N_{+\ell} = \sum_{j=1}^J N_{j\ell}$, are known for all j and ℓ from external data. For example, $X_1 = \text{sex}$, $X_2 = \text{race}$, where the marginal distributions of sex and race are available, but the joint distribution in the sex by race table is not.

Raking estimates $\{N_{j\ell}^*\}$ of $\{N_{j\ell}\}$ satisfy the marginal constraints

$$N_{j+}^* = \sum_{\ell=1}^L N_{j\ell}^* = N_{j+}, \quad j = 1, \dots, J, \quad N_{+\ell}^* = \sum_{j=1}^J N_{j\ell}^* = N_{+\ell}, \quad \ell = 1, \dots, L$$

and have the form

$$N_{j\ell}^* = a_j b_\ell n_{j\ell}, \quad j = 1, \dots, J, \quad \ell = 1, \dots, L,$$

for certain row constants $\{a_j, j = 1, \dots, J\}$ and column constants $\{b_\ell, \ell = 1, \dots, L\}$. The $\{N_{j\ell}^*\}$ table has margins that equal the known margins $\{N_{j+}\}$ and $\{N_{+\ell}\}$ and two-way associations that equal those in the $\{n_{j\ell}\}$ table. The raked class counts $\{N_{j\ell}^*\}$ can be calculated by an iterative proportional fitting procedure (Bishop et al. 1975), where current estimates are scaled by row or column factors to match the marginal totals $\{N_{j+}\}$ and $\{N_{+\ell}\}$, respectively. That is, at the first step, the estimates

$$N_{j\ell}^{(1)} = n_{j\ell} (N_{j+}/n_{j+}),$$

which match the row marginals, $\{N_{j+}\}$, are calculated. Then, estimates

$$N_{j\ell}^{(2)} = N_{j\ell}^{(1)} \left(N_{+\ell}/N_{+\ell}^{(1)} \right),$$

that match the column marginals $\{N_{+\ell}\}$ are constructed. Then,

$$N_{j\ell}^{(3)} = N_{j\ell}^{(2)} \left(N_{j+}/N_{j+}^{(2)} \right),$$

and so on, until convergence. Convergence and statistical properties of this procedure are discussed by Ireland and Kullback (1968), who show, in particular, that the raked estimates $\{N_{j\ell}^*/N\}$ of the class proportions are efficient asymptotically normal estimates under a multinomial assumption for the class counts $\{n_{j\ell}\}$ and, as such, are asymptotically equivalent to the (harder to calculate) ML estimates under the multinomial model. This algorithm is a special case of the Gauss–Seidel method for iterative maximization.

Combining the raked sample counts $\{N_{j\ell}^*\}$ with the respondent means $\{\bar{y}_{j\ell}\}$ yields the raked estimate of \bar{Y} :

$$\bar{y}_{\text{rake}} = \frac{1}{N} \sum_{j=1}^J \sum_{\ell=1}^L N_{j\ell}^* \bar{y}_{j\ell R}, \quad (3.16)$$

which might be expected to have sampling variance properties somewhere between \bar{y}_{wc} and \bar{y}_{ps} . Note that this estimate is not defined when $r_{j\ell} = 0, n_{j\ell} \neq 0$ for some j, ℓ , and in this situation, some other estimate of the mean for that class is required. See Little (1993b) for further discussion of alternatives.

3.3.3 Inference from Weighted Data

Weighted CC estimates are often relatively simple to compute, but the computation of appropriate standard errors, even asymptotically, is less straightforward. For simple settings such as weighting class adjustment for simple random sampling, formulae are available for estimating the standard errors. For more complex situations, methods based on Taylor Series expansions (Robins

et al. 1995), balanced repeated replication or jackknifing can be applied. Statistical packages are available for computing asymptotic standard errors of estimates from complex sample survey designs that include weighting, clustering, and stratification. However, these programs typically treat the weights as fixed and known, whereas nonresponse weights are computed from the observed data and hence are themselves subject to sampling uncertainty. The practical impact on the standard errors from ignoring this source of variability is unclear. Valliant (2004) finds that ignoring sampling variability in the weights leads to undercoverage, particularly in small samples, whereas jackknifing tends to yield conservative intervals. A computationally intensive approach that yields valid asymptotic inferences is to apply a sample reuse method, such as balanced repeated replication or jackknifing to the sample, and recalculate the weights separately for each replicate or bootstrap sample. Subasymptotic performance of these approaches is relatively unstudied.

3.3.4 Summary of Weighting Methods

Weighting is a relatively simple device, both conceptually and computationally, for reducing bias from CC analysis. The methods are simple in that they yield the same weight for all variables measured for each unit. Because the methods discard the incomplete units and do not provide an automatic control of sampling variance, they are most useful when covariate information is limited and the sample size is large so that bias is a more serious issue than sampling variance.

3.4 Available-Case Analysis

CC analysis is typically wasteful for univariate analyses, such as estimation of means and marginal frequency distributions, because values of all observed variables in incomplete units are discarded. The loss in efficiency can be particularly large for data sets involving a large number of variables. For example, if there are 20 variables, and each variable independently has a 10% chance of being missing, then the expected proportion of complete units is $0.9^{20} \doteq 0.12$. That is, only about $12/0.9=13\%$ of the observed data values will be retained.

A natural alternative procedure for univariate analyses is to include all units where that variable is present, an option that has been termed *available-case* (AC) analysis. The method uses all the available values; its disadvantage is that the sample changes from variable to variable according to the pattern of missing data. This variability complicates checks for whether tables computed for various conceptual sample bases (e.g., all women, ever-married women, currently married women in a demographic fertility survey) are correctly defined. It also creates problems of comparability across variables if the missing data mechanism is a function of the variables under study, that is, if it is not MCAR.

Under MCAR, estimates of means and variances can be calculated using the AC procedure we have described, but extensions are required to estimate measures of covariation, such as covariances or correlations. A natural extension is *pairwise* AC methods, where measures of covariation for Y_j and Y_k are based on units i for which both y_{ij} and y_{ik} are observed. In particular, one can compute pairwise covariances:

$$s_{jk}^{(jk)} = \sum_{i \in I_{jk}} (y_{ij} - \bar{y}_j^{(jk)}) (y_{ik} - \bar{y}_k^{(jk)}) / (n^{(jk)} - 1), \quad (3.17)$$

where I_{jk} is the set of $n^{(jk)}$ units with both Y_j and Y_k observed, and the means $\bar{y}_j^{(jk)}, \bar{y}_k^{(jk)}$ are calculated over that set of units. Let $s_{jj}^{(j)}$ denote the sample variance of Y_j over the set I_j of units with Y_j observed. Combining these variance estimates with the covariances estimated from (3.17) yields the following estimate of the correlation:

$$r_{jk}^* = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}. \quad (3.18)$$

A criticism of (3.18) is that, unlike the population correlation being estimated, r_{jk}^* can lie outside the range $(-1, 1)$. This difficulty is avoided by computing pairwise correlations, where variances are estimated from the same sample base as the covariance:

$$r_{jk}^{(jk)} = s_{jk}^{(jk)} / \sqrt{s_{jj}^{(jk)} s_{kk}^{(jk)}}. \quad (3.19)$$

This estimate is discussed by Matthai (1951). It corresponds to the covariance estimate

$$s_{jk}^* = r_{jk}^{(jk)} \sqrt{s_{jj}^{(j)} s_{kk}^{(k)}}. \quad (3.20)$$

Still more estimates can be constructed by replacing the means $\bar{y}_j^{(jk)}$ in (3.17)–(3.20) by estimates $\bar{y}_j^{(j)}$ from all units with Y_j observed. Applying this idea to (3.17) yields

$$\tilde{s}_{jk}^{(jk)} = \sum_{i \in I_{jk}} (y_{ij} - \bar{y}_j^{(j)}) (y_{ik} - \bar{y}_k^{(k)}) / (n^{(jk)} - 1) \quad (3.21)$$

an estimate originally discussed in Wilks (1932).

Pairwise AC estimates such as (3.17)–(3.21) attempt to recover information in partially recorded units that is lost by CC analysis. Under MCAR, Eqs. (3.17)–(3.21) yield consistent estimates of the covariances and correlations

being estimated. When considered collectively, however, all the estimates have deficiencies that can severely limit their utility in practical problems.

For example (3.18) can yield correlations outside the acceptable range. On the other hand (3.19) yields correlations that always lie between ± 1 . For $K > 3$ variables, both (3.18) and (3.19) can yield estimated correlation matrices that are not positive definite. To take an extreme artificial example, consider the following data set with 12 observations on 3 variables (? denotes missing):

Y_1	1	2	3	4	1	2	3	4	?	?	?	?
Y_2	1	2	3	4	?	?	?	?	1	2	3	4
Y_3	?	?	?	?	1	2	3	4	4	3	2	1

Equation (3.19) yields $r_{12}^{(12)} = 1$, $r_{13}^{(13)} = 1$, $r_{23}^{(23)} = -1$. These estimates are clearly unsatisfactory because $\text{Corr}(Y_1, Y_2) = \text{Corr}(Y_1, Y_3)$ implies $\text{Corr}(Y_2, Y_3) = 1$, not -1 . In the same way, covariance matrices based on (3.17) or (3.20) are not necessarily positive definite. Because many analyses based on the covariance matrix, including multiple regression, require a positive-definite matrix, ad hoc modifications are required for these methods when this condition is not satisfied. Any method that can produce parameter estimates outside the parameter space is not satisfactory.

Because AC methods *apparently* make use of all the data, one might expect them to be more efficient than CC methods, a conclusion supported in simulations by Kim and Curry (1977) when the data are MCAR, and correlations are modest. Other simulations, however, indicate superiority for CC analysis when correlations are large (Haitovsky 1968; Azen and Van Guilder 1981). Neither method, however, is generally satisfactory. Although AC estimates are easy to compute, asymptotic standard errors are more complex (Van Praag et al. 1985).

Problems

- 3.1 List some standard multivariate statistical analyses that are based on sample means, variances, and correlations.
 - 3.2 Show that if missingness (of Y_1 or Y_2) depends only on Y_2 , and Y_1 has a linear regression on Y_2 , then the sample regression of Y_1 on Y_2 based on complete units yields unbiased estimates of the regression parameters.
 - 3.3 Show that for dichotomous Y_1 and Y_2 , the odds ratio based on complete units is a consistent estimate of the population odds ratio if the logarithm of the probability of response is an additive function of Y_1 and Y_2 .
- Data for Problems 3.4–3.6: a simple random sample of 100 individuals in a county is interviewed for a health survey, yielding the following data:

Cholesterol				
Age group	Sample size	Number of respondents	Mean	SD
20–30	25	22	220	30
30–40	35	27	225	35
40–50	28	16	250	44
50–60	12	5	270	41

- 3.4** Compute the mean cholesterol for the respondent sample and its standard error. Assuming normality, compute a 95% confidence interval for the mean cholesterol for respondents in the county. Should this interval be applied to summarize all individuals in the county?
- 3.5** Compute the weighting class estimate (3.6) of the mean cholesterol level in the population and its estimated mean squared error (3.7). Thereby construct an approximate 95% confidence interval for the population mean and compare it with the result of Problem 3.4. For each procedure, what assumptions are made about the missingness mechanism?
- 3.6** Suppose census data yield the following age distribution for the county of interest in Problems 3.4 and 3.5: 20–30: 20%; 30–40: 40%; 40–50: 30%; and 50–60: 10%. Calculate the poststratified estimate of mean cholesterol, its associated standard error, and a 95% confidence interval for the population mean.
- 3.7** Calculate Horvitz–Thompson and weighting class estimates of the overall mean of Y in the following artificial example of a stratified random sample, where the x_i and y_i values displayed are observed, the selection probabilities π_i are known, and the response probabilities, ϕ_i , are (unrealistically) assumed to be known for the Horvitz–Thompson estimate but are assumed to be unknown for the weighting class estimates. Note that various weighting class estimates could be created, depending on how the weighting classes are defined.

Data for Problem 3.7:

x_i	1	2	3	4	5	6	7	8	9	10
y_i	1	4	3	2	6	10	14	?	?	?
π_i	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5	0.5
ϕ_i	1	1	1	0.9	0.9	0.8	0.7	0.6	0.5	0.1

- 3.8** Apply the Cassel et al. (1983) estimate, discussed in Example 3.7, to the data of Problem 3.7. Comment on the resulting weights as compared with those of the weighting class estimate.

- 3.9** The following table shows respondent means of an incomplete variable Y (income in \$1000), and response rates (respondent sample size/sample size), classified by three fully observed covariates: Age (<30 , >30), marital status (single, married), and sex (male, female). Note that weighting classes cannot be based on age, marital status, and sex because there is one class with four units sampled, none of whom responded.
Respondent means and response rates, classified by age, marital status, and gender:

Gender		Male		Female	
Marital status		Single	Married	Single	Married
Age					
<30	Single	20.0	21.0	16.0	16.0
	Married	24/25	5/16	11/12	2/4
>30	Single	30.0	36.0	18.0	?
	Married	15/20	2/10	8/12	0/4

Calculate the following estimates of the mean of Y , both for the entire population and for the subpopulation of males:

- (a) The unadjusted mean based on complete units.
- (b) The weighted mean from response propensity stratification, with three strata defined by combining classes in the table with response rates less than 0.4, between 0.4 and 0.8, and greater than 0.8.
- (c) The mean from mean imputation within adjustment classes defined as in (b). Explain why adjusted estimates are higher than the unadjusted estimates.

- 3.10** Generalize the response propensity method in Example 3.7 to a monotone pattern of missing data (see Little 1986; Robins et al. 1995).
- 3.11** Oh and Scheuren (1983, section 4.4.3) propose an alternative to the raked estimate \bar{y}_{rake} in Eq. (3.16), where the estimated counts $N_{j\ell}^*$ are found by raking the respondent sample sizes $\{r_{j\ell}\}$ instead of $\{n_{j\ell}\}$. Show that (i) unlike \bar{y}_{rake} , this estimate exists when $r_{j\ell} = 0, n_{j\ell} \neq 0$ for some j and ℓ , and (ii) the estimate is biased unless the expectation of $r_{j\ell}/n_{j\ell}$ can be written as a product of row and column effects.

- 3.12** Show that raking the class sample sizes and raking the class respondent sample sizes (as in Problem 3.11) yield the same estimate if and only if

$$p_{ij}p_{kl}/p_{il}p_{jk} = 1 \quad \text{for all } i, j, k, \text{ and } l,$$

where p_{ij} is the response rate in class (i, j) of the table.

- 3.13** Compute raked estimates of the class counts from the sample counts and respondent counts in (a) and (b) below, using population marginal counts in (c):

(a) Sample $\{n_{jj}\}$			(b) Respondent $\{r_{jl}\}$			(c) Population $\{N_{jl}\}$		
8	10	18	5	9	14	?	?	300
15	17	32	5	8	13	?	?	700
23	27	50	10	17	27	500	500	1000

- 3.14** For the data in Problem 3.13, compute the odds ratio of response rates discussed in Problem 3.12. Repeat the computation with the respondent counts 5 and 8 in the second row of (b) in Problem 3.13 interchanged. By comparing these odds ratios, predict which set of the raked respondent counts will be closer to the raked sample counts. Then compute the raked counts for (b) with the modified second row and check your prediction.
- 3.15** Construct a data set where the estimated correlation given by Eq. (3.18) lies outside the range $(-1, 1)$.
- 3.16** (a) Why does the estimated correlation (3.19) always lie in the range $(-1, 1)$? (b) Suppose the means $\bar{y}_j^{(jk)}$ and $\bar{y}_k^{(jk)}$ in the definitions of $s_{jj}^{(jk)}, s_{jk}^{(jk)}$, and $s_{kk}^{(jk)}$ in (3.17) are replaced by the means $\bar{y}_j^{(j)}$ and $\bar{y}_k^{(k)}$ from all available units for those variables. Is the resulting estimated correlation (3.19) always in the range $(-1, 1)$? Either prove that it is or provide a counterexample.
- 3.17** When the data are not MCAR, consider the relative merits of CC analysis and AC analysis for estimating (a) means, (b) correlations, and (c) regression coefficients.
- 3.18** Review the results of Haitovsky (1968), Kim and Curry (1977), and Azen and Van Gulder (1981). Describe situations where CC analysis is more sensible than AC analysis, and vice versa.

4

Single Imputation Methods

4.1 Introduction

Both complete-case and available-case analyses make no use of units with Y_j missing when estimating either the marginal distribution of Y_j or measures of covariation between Y_j and other variables. Intuitively, this is a mistake. Suppose a unit with Y_j (e.g., height) missing has the value of another variable Y_k (e.g., weight) that is highly correlated with Y_j . It is tempting to predict the missing value of Y_j from Y_k and then to include the filled-in (or *imputed*) value in analyses involving Y_j . We now discuss methods that impute (that is fill in) the values of variables that are missing. These methods can be applied to impute one value for each missing variable (single imputation), or to impute more than one value (multiple imputation), to allow appropriate assessment of imputation uncertainty.

Imputation is a general and flexible method for handling missing data problems. However, it has pitfalls. In the words of Dempster and Rubin (1983):

The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.

Imputations should be conceptualized as draws from a predictive distribution of the missing values and require a method for creating a predictive distribution for the imputation based on the observed data. There are two generic approaches to generating this distribution:

Explicit modeling: The predictive distribution is based on a formal statistical model (e.g., multivariate normal), and hence, the assumptions are explicit.

These approaches are introduced in this chapter (see Examples 4.2–4.5) but receive systematic attention in Parts II and III of the book.

Implicit modeling: The focus is on an algorithm, which may imply an underlying model; assumptions are implicit, but they still need to be carefully assessed to ensure that they are reasonable. These approaches are discussed in Examples 4.8–4.12.

Explicit modeling methods include the following:

- (a) *Mean imputation:* Where means from the responding units in the sample are substituted. The means may be formed within cells or classes analogous to the weighting classes discussed in Chapter 3. Mean imputation then leads to estimates similar to those found by weighting, provided the sampling weights are constant within weighting classes.
- (b) *Regression imputation:* Replaces missing values for a unit by their predicted values from a regression of the missing variable on variables observed for the unit, usually calculated from units with both variables observed. Mean imputation can be regarded as a special case of regression imputation where the predictor variables are dummy indicator variables for the cells within which the means are imputed.
- (c) *Stochastic regression imputation:* Replaces missing values by a value predicted by regression imputation plus a randomly selected residual, drawn to reflect uncertainty in the predicted value. With normal linear regression models, the residual will naturally be normal with zero mean and variance equal to the residual variance in the regression. With a binary outcome, as in logistic regression, the predicted value is a probability of 1 vs. 0, and the imputed value is a 1 or 0 drawn with that probability. Herzog and Rubin (1983) describe a two-stage procedure that uses stochastic regression for both normal and binary outcomes.

Implicit modeling methods include the following:

- (d) *Hot deck imputation:* Which imputes individual values drawn from “similar” responding units. Hot deck imputation is common in survey practice and can involve very elaborate schemes for selecting units for imputation that are similar. See, for example Ernst (1980), Kalton and Kish (1981), Ford (1983), Sande (1983), David et al. (1986), Marker et al. (2002), and Andridge and Little (2010).
- (e) *Substitution:* A method for dealing with unit nonresponse at the fieldwork stage of a survey, replaces nonresponding units with alternative units not yet selected into the sample. For example, if a household cannot be contacted, then a previously nonselected household in the same housing block may be substituted. The tendency to treat the resulting sample as complete should be resisted because the substituted units are respondents and

hence may differ systematically from nonrespondents. Hence, at the analysis stage, substituted values should be regarded as imputed values for responses that are not known.

- (f) *Cold deck imputation*: Replaces a missing value of a variable by a constant value from an external source, such as a value from a previous survey. As with substitution, current practice usually treats the resulting data as a complete sample, that is ignores the consequences of imputation. Satisfactory theory for the analysis of data obtained by cold deck imputation is either obvious or lacking.
- (g) *Composite methods*: Can also be defined as methods that combine ideas from different methods. For example, hot deck and regression imputation can be combined by calculating predicted means from a regression but then adding a residual randomly chosen from the empirical residuals to the predicted value when forming values for imputation. See, for example, Schieber (1978) and David et al. (1986).

We now discuss these methods in more detail. An important limitation of single imputation methods described here is that standard sampling variance formulas applied to the filled-in data systematically underestimate the true sampling variance of estimates, even if the model used to generate the imputations is correct. Methods that allow valid estimates of the sampling variance of estimates to be calculated using standard complete-data procedures are introduced in Chapter 5 and further discussed in Chapters 9 and 10 in the context of model-based methods.

4.2 Imputing Means from a Predictive Distribution

4.2.1 Unconditional Mean Imputation

Let y_{ij} be the value of Y_j for unit i . A simple form of imputation is to estimate missing values y_{ij} by $\bar{y}_j^{(j)}$, the mean of the recorded values of Y_j . The average of the observed and imputed values is then clearly $\bar{y}_j^{(j)}$, the estimate from available-case analysis. The sample variance of the observed and imputed values is $s_{jj}^{(j)}(n^{(j)} - 1)/(n - 1)$, where $s_{jj}^{(j)}$ is the estimated variance from the $n^{(j)}$ available units. Under missing completely at random (MCAR), $s_{jj}^{(j)}$ is a consistent estimate of the true variance, so the sample variance from the filled-in data set systematically underestimates the true variance by a factor of $(n^{(j)} - 1)/(n - 1)$. This underestimation is a natural consequence of imputing missing values at the center of the distribution. Imputation distorts the empirical distribution of the sampled Y -values, as well as quantities that are not linear in the data, such as variances, percentiles, or measures of shape, which are not estimated

consistently using standard complete-data methods applied to the filled-in data. A bias occurs if the values of Y_j are grouped into subclasses for cross-tabulation because missing values in an adjustment cell are all replaced by a common mean value and hence are classified in the same subclass of Y_j .

The sample covariance of Y_j and Y_k from the filled-in data is $\tilde{s}_{jk}^{(jk)}(n^{(jk)} - 1)/(n - 1)$, where $n^{(jk)}$ is the number of units with Y_j and Y_k observed and $\tilde{s}_{jk}^{(jk)}$ is given by Eq. (3.21). Under MCAR, $\tilde{s}_{jk}^{(jk)}$ is a consistent estimate of the covariance, so the estimated covariance from the filled-in data underestimates the magnitude of the covariance by a factor $(n^{(jk)} - 1)/(n - 1)$. Thus, although the covariance matrix from the filled-in data is positive semidefinite, the variances and covariances are systematically attenuated. Obvious adjustment factors, namely, $(n - 1)/(n^{(j)} - 1)$ for the variance of Y_j and $(n - 1)/(n^{(jk)} - 1)$ for the covariance of Y_j and Y_k , simply yield available-case estimates $\tilde{s}_{jk}^{(jk)}$ for the covariances and $s_{jj}^{(j)}$ for the variances. As noted in Section 3.4, the resulting covariance matrix is not generally positive definite and tends to be unsatisfactory, particularly when the variables are highly correlated.

Considering the biases resulting from imputing the unconditional mean, the method cannot be recommended.

4.2.2 Conditional Mean Imputation

An improvement over unconditional mean imputation imputes conditional means given observed values. We consider two examples of this approach.

Example 4.1 *Imputing Means Within Adjustment Cells.* A common method in surveys is to classify nonrespondents and respondents into J adjustment classes, analogous to weighting classes, based on the observed variables, and then impute the respondent mean for nonrespondents in the same adjustment class. Assume equal probability sampling with constant sample weights, and let \bar{y}_{jR} be the respondent mean for a variable Y in class j . The resulting estimate of the mean of Y from the filled-in data is

$$\frac{1}{n} \sum_{j=1}^J \left(\sum_{i=1}^{r_j} y_{ij} + \sum_{i=r_j+1}^{n_j} \bar{y}_{jR} \right) = \frac{1}{n} \sum_{j=1}^J n_j \bar{y}_{jR} = \bar{y}_{wc},$$

the estimator (3.6) that weights by the inverse of the proportion of respondents in each class. If the proportions of the population in each class are known from external data, then the poststratified estimator \bar{y}_{ps} , Eq. (3.10), can also be derived as an estimator based on mean imputation. For more on the relationship between imputation and weighting, see Oh and Scheuren (1983), David et al. (1983), and Little (1986).

Example 4.2 Regression Imputation. Consider univariate nonresponse, with Y_1, \dots, Y_{K-1} fully observed and Y_K observed for the first r units and missing for the last $n - r$ units. Regression imputation computes the regression of Y_K on Y_1, \dots, Y_{K-1} based on the r complete units and then fills in the missing values as predictions from the regression, an approach similar to that in Chapter 2. Specifically, suppose unit i has y_{iK} missing and $y_{i1}, \dots, y_{i,K-1}$ observed. The missing value is then imputed using the regression equation:

$$\hat{y}_{iK} = \tilde{\beta}_{K0 \cdot 12 \dots K-1} + \sum_{j=1}^{K-1} \tilde{\beta}_{Kj \cdot 12 \dots K-1} y_{ij}, \quad (4.1)$$

where $\tilde{\beta}_{K0 \cdot 12 \dots K-1}$ is the intercept and $\tilde{\beta}_{Kj \cdot 12 \dots K-1}$ is the coefficient of Y_j in the regression of Y_K on Y_1, \dots, Y_{K-1} based on the r complete units. If the observed variables are indicator variables for a categorical variable, the predictions (4.1) are respondent means within classes defined by that variable, and the method reduces to that of Example 4.1. More generally, the regression might include continuous and categorical variables, interactions, and less restrictive parametric forms, such as splines, might be substituted to improve the predictions.

Regression imputation is illustrated graphically for $K = 2$ variables in Figure 4.1. The points marked as pluses represent units with Y_1 and Y_2 both observed. These points are used to calculate the least squares regression line of Y_2 on Y_1 , $\hat{y}_{i2} = \tilde{\beta}_{20 \cdot 1} + \tilde{\beta}_{21 \cdot 1} y_{i1}$. Units with Y_1 observed but Y_2 missing are represented by circles on the Y_1 -axis. Regression imputation replaces them by the dots lying on the regression line. Missing values of Y_1 for units with Y_2 observed would be imputed on the regression line of Y_1 on Y_2 , the other line in the diagram.

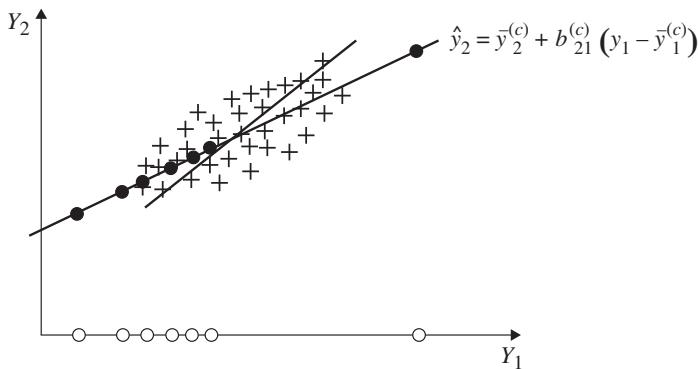


Figure 4.1 Example 4.2: regression imputation for $K = 2$ variables.

Example 4.3 Buck's Method. Buck's method (Buck 1960) extends regression imputation to a general pattern of missing values, for the situation where missing variables have linear regressions on the observed variables. The method first estimates the mean μ and covariance matrix Σ from the sample mean and covariance matrix of the complete units and then uses these estimates to calculate the least squares linear regressions of the missing variables on the observed variables for each missing data pattern. Predictions of the missing values for each observation are obtained by substituting the values of the present variables in the regressions. The computation of different linear regressions for the set of units with each pattern of missing data is simple using the sweep operator discussed in Section 7.4.3.

The averages of observed and imputed values from this procedure are consistent estimates of the means under MCAR and mild assumptions about the moments of the distribution (Buck 1960). They are also consistent when the missingness mechanism depends on variables that are observed, although additional assumptions are required in this case. In particular, suppose that for the data in Figure 4.1, missingness of Y_2 depends on the values of Y_1 , so that missing at random (MAR) holds, even though the distribution of Y_1 for complete and incomplete units is different. Buck's method projects the incomplete units to the regression line, a process that makes the assumption that the regression of Y_2 on Y_1 is linear. This assumption is particularly tenuous if the imputation involves extrapolation beyond the range of the complete data, as occurs for the incomplete units with the two smallest and the single largest Y_1 values in Figure 4.1.

The filled-in data from Buck's method yield reasonable estimates of means, particularly if the joint normality assumption appears plausible. The sample variances and covariances from the filled-in are biased, although the bias is less than that obtained when unconditional means are substituted. Specifically, the sample variance of Y_j from the filled-in data underestimates σ_{jj} by $(n - 1)^{-1} \sum_{i=1}^n \sigma_{jj \cdot \text{obs}, i}$, where $\sigma_{jj \cdot \text{obs}, i}$ is the residual variance from regressing Y_j on the variables observed in unit i , if y_{ij} is missing, and is zero if y_{ij} is observed. The sample covariance of Y_j and Y_k has a bias of $(n - 1)^{-1} \sum_{i=1}^n \sigma_{jk \cdot \text{obs}, i}$ where $\sigma_{jk \cdot \text{obs}, i}$ is the residual covariance of Y_j and Y_k from the multivariate regression of Y_j , Y_k on the variables observed for unit i , if both y_{ij} and y_{ik} are missing, and zero otherwise. A consistent estimate of Σ can be constructed under the MCAR assumption by substituting consistent estimates of $\sigma_{jj \cdot \text{obs}, i}$ and $\sigma_{jk \cdot \text{obs}, i}$ (such as estimates based on the sample covariance matrix of the complete units, sample sizes permitting) in the expressions for bias and then adding the resulting quantities to the sample covariance matrix of the filled-in data. This method is closely related to a single iteration of the maximum likelihood procedure presented in Section 11.2 and can be viewed as a historical precursor of that method.

4.3 Imputing Draws from a Predictive Distribution

4.3.1 Draws Based on Explicit Models

We have seen that sample variances and covariances based on the filled-in data are distorted by mean imputation. The distortion of the marginal distribution from mean imputation also leads to bias when the tails of the distribution are being studied. For example, an imputation method that imputes conditional means for missing incomes tends to underestimate the percentage of units in poverty. Because such “best prediction” imputations systematically underestimate variability, standard errors calculated from the filled-in data are too small, leading to invalid inferences.

These considerations suggest an alternative strategy where imputations are random draws from a predictive distribution of plausible values of the missing value or set of values, rather than from the center of this distribution. As before, it is important to condition on the observed data when creating the predictive distribution, as in the following example.

Example 4.4 Stochastic Regression Imputation. Consider the data of Example 4.2, but suppose that instead of imputing the conditional mean (4.1), we impute a conditional *draw*:

$$\hat{y}_{iK} = \tilde{\beta}_{K0 \cdot 12 \dots K-1} + \sum_{j=1}^{K-1} \tilde{\beta}_{Kj \cdot 12 \dots K-1} y_{ij} + z_{iK}, \quad (4.2)$$

where z_{iK} is a random normal deviate with mean 0 and variance $\tilde{\sigma}_{KK \cdot 12 \dots K-1}$, the residual variance from the regression of Y_K on Y_1, \dots, Y_{K-1} based on the complete units. The addition of the random normal deviate makes the imputation a draw from the predictive distribution of the missing values, rather than the mean. As a result, the distortions from imputing the mean of the predictive distributions are ameliorated, as the following summary example illustrates.

Example 4.5 Comparison of Methods for Bivariate Monotone MCAR Data. The advantages of imputing conditional draws can be illustrated for the case of bivariate normal monotone missing data with Y_1 fully observed, Y_2 missing for a fraction $\lambda = (n - r)/n$ units, and a MCAR mechanism. Table 4.1 shows the large sample bias (that is the expectation of the estimate from the filled-in data minus the true value, ignoring order $1/n$ terms) of four parameters, namely the mean and variance of Y_2 , the regression coefficient of Y_2 on Y_1 , and the regression coefficient of Y_1 on Y_2 , when standard least squares estimates are computed using the filled-in data. Four imputation methods are applied to Y_2 , namely:

- (1) *Umean*: Unconditional means, where the respondent mean \bar{y}_{2R} is imputed for each missing value of Y_2 .

Table 4.1 Example 4.5: bivariate normal monotone MCAR data: large sample bias of four imputation methods

Method	Parameter			
	μ_2	σ_{22}	$\beta_{21 \cdot 1}$	$\beta_{12 \cdot 2}$
Umean	0 ^a	$-\lambda\sigma_{22}$	$-\lambda\beta_{21 \cdot 1}$	0 ^a
Udraw	0	0	$-\lambda\beta_{21 \cdot 1}$	$-\lambda\beta_{12 \cdot 2}$
Cmean	0	$-\lambda(1 - \rho^2)\sigma_{22}$	0 ^a	$\frac{\lambda(1 - \rho^2)}{1 - \lambda(1 - \rho^2)}\beta_{12 \cdot 2}$
Cdraw	0	0	0	0

λ = fraction of missing data.

^aIndicates estimator is same as CC estimate.

- (2) *Udraw*: Unconditional draws, where a random normal deviate with mean 0 and variance $\tilde{\sigma}_{22}$ is added to \bar{y}_{2R} . Here, $\tilde{\sigma}_{22}$ is the sample variance of Y_2 based on the complete units.
- (3) *Cmean*: Conditional means, as discussed in Example 4.2 with $K = 2$.
- (4) *Cdraw*: Conditional draws, as discussed in Example 4.4 with $K = 2$.

Table 4.1 shows that under MCAR, all four imputation methods yield consistent estimates of μ_2 , but Umean and Cmean both underestimate the variances, and Udraw leads to attenuation of the regression coefficients. Only Cdraw yields consistent estimates of all the parameters from the filled-in data. This result concerning Cdraw also holds under the less restrictive MAR assumption.

Cdraw is the generally preferred imputation method in this example, but it has two drawbacks. First, the random draws added to the conditional mean imputations entail a loss of efficiency. Specifically, the large sample variance of the Cdraw estimate of μ_2 can be shown to be $[1 - \lambda\rho^2 + (1 - \rho^2)\lambda(1 - \lambda)]\sigma_{22}/r$, which is larger than the large sample sampling variance of the Cmean estimate of μ_2 , namely $[1 - \lambda\rho^2]\sigma_{22}/r$. Second, the standard errors of the Cdraw parameter estimates from the filled-in data are too small because they do not incorporate imputation uncertainty. Multiple imputation, which we discuss in Chapters 5 and 10, addresses both of these deficiencies of the Cdraw method.

Example 4.6 Missing Covariates in Regression. The last column in Table 4.1 is the simplest form of the problem of missing covariates in regression. When units involve covariates that are missing and covariates that are observed, it

is common practice to condition on the observed covariates when imputing the missing covariates. A commonly asked question in this setting is whether imputations of the missing covariates should also condition on the outcome Y . It may appear circular to condition imputations on Y when the final objective is to regress Y on the full set of covariates, and conditioning on Y does lead to bias when conditional means are imputed. However, our recommended approach is to impute draws (not means) from the conditional distribution of the missing covariates given the observed covariates and Y ; this approach yields consistent estimates of the regression coefficients and hence is not circular. The traditional approach of imputing means that condition on the observed covariates but not Y also yields consistent estimates of the regression coefficients under certain conditions but that method yields estimates of regression coefficients that can be less efficient estimates than complete-case analysis, and yields inconsistent estimates of other parameters (e.g., variances, correlations). See Little (1992) for further discussion.

Example 4.7 Regression Calibration for Measurement Error in Regression. In Example 1.15, we described measurement error as a missing data problem. Regression imputation is a common method in the analysis of internal calibration data (see Figure 1.4a). The analysis concerns the regression of Y on X and Z . The main study data are a random sample on Y , Z , and W , where X is unobserved, and W is a proxy for X measured with error. Information relating W and X is obtained from a calibration sample, which for internal calibration data includes measurements on W and X , as well as on Y , Z . If the calibration sample is a random subsample from the main study, the missingness of X is MCAR.

Regression calibration (RC, Carroll and Stefanski 1990) estimates the regression of X on W and Z using the calibration data, and then substitutes the unknown values X in the main study with mean predictions $\hat{X}_{\text{RC}} = E(X | W, Z)$ from this regression. The RC estimates are then obtained by regressing Y on \hat{X}_{RC} and Z . The regression imputations from this RC method yield consistent estimates of the regression of Y on X , W , and Z . It yields consistent estimates of the regression of Y on X and Z under the assumption that measurement error is nondifferential, which means that W is independent of X given Y and Z because in that case $E(Y|X, Z) = E(Y|X, Z, W)$. If this assumption is violated, the RC estimates are generally biased.

Two estimates of the regression coefficient γ_X are available here, the RC estimate $\hat{\gamma}_{X,\text{RC}}$ and the least squares estimate $\hat{\gamma}_{X,\text{LSCalib}}$, obtained from fitting the linear regression model to the calibration sample data on (Y, X, Z) . Efficient regression calibration (ERC, Spiegelman et al. 2001) combines these two estimates:

$$\hat{\gamma}_{X,\text{ERC}} = w_{\text{RC}} \hat{\gamma}_{X,\text{RC}} + (1 - w_{\text{RC}}) \hat{\gamma}_{X,\text{LSCalib}},$$

where $w_{RC} = \widehat{\text{var}}(\hat{Y}_{X,RC})^{-1} [\widehat{\text{var}}(\hat{Y}_{X,RC})^{-1} + \widehat{\text{var}}(\hat{Y}_{X,LSCalib})^{-1}]^{-1}$ and $\widehat{\text{var}}(\hat{Y}_{X,RC})$ and $\widehat{\text{var}}(\hat{Y}_{X,LSCalib})$ are the estimated sampling variances of $\hat{Y}_{X,RC}$ and $\hat{Y}_{X,LSCalib}$, respectively. The sampling variance of $\hat{Y}_{X,ERC}$ is computed approximately as $\widehat{\text{var}}(\hat{Y}_{X,ERC}) = [\widehat{\text{var}}(\hat{Y}_{X,RC})^{-1} + \widehat{\text{var}}(\hat{Y}_{X,LSCalib})^{-1}]^{-1}$. ERC is more efficient than RC, particularly when the calibration data set is large.

The standard error of the RC and ERC estimates can be estimated using asymptotic calculations, or by bootstrapping the main and calibration samples (as discussed in Chapter 5).

4.3.2 Draws Based on Implicit Models – Hot Deck Methods

Hot deck procedures impute missing values using observed values from similar responding units in the sample. The hot deck literally refers to the deck of matching Hollerith cards for the donors available for a nonrespondent. Suppose, as before, that a sample of n out of N units is selected, and r out of the n sampled values of a variable Y are recorded, where n , N , and r are treated throughout this section as fixed. For simplicity, we label the first n units, $i = 1, \dots, n$ as sampled, and the first $r < n$ units as respondents. Given an equal probability sampling scheme, the mean Y may be estimated from the filled-in data as the mean of the responding and the imputed units:

$$\bar{y}_{HD} = \{r\bar{y}_R + (n - r)\bar{y}_{NR}^*\} / n, \quad (4.3)$$

where \bar{y}_R is the mean of the respondent units, and

$$\bar{y}_{NR}^* = \sum_{i=1}^r \frac{H_i y_i}{n - r},$$

where H_i is the number of times y_i is used as a substitute for a missing value of Y , with $\sum_{i=1}^r H_i = n - r$, the number of missing units. The properties of \bar{y}_{HD} depend on the procedure used to generate the numbers $\{H_1, \dots, H_r\}$. The simplest theory is obtained when imputed values can be regarded as selected from the values for the responding units by a probability sampling design, so that the distribution of $\{H_1, \dots, H_r\}$ in repeated applications of the hot deck method is known. The mean and sampling variance of \bar{y}_{HD} can then be written as follows:

$$E(\bar{y}_{HD}) = E(E(\bar{y}_{HD} | Y_{(0)})), \quad (4.4)$$

$$\text{Var}(\bar{y}_{HD}) = \text{Var}(E(\bar{y}_{HD} | Y_{(0)})) + E(\text{Var}(\bar{y}_{HD} | Y_{(0)})), \quad (4.5)$$

where the inner expectations and variances are over the distribution of $\{H_1, \dots, H_r\}$ given the observed data $Y_{(0)}$, and the outer expectations and variances are over the model distribution of Y , or the distribution of the sampling indicators

for design-based inference (see Example 3.5). The second term in (4.5) represents the added sampling variance from the stochastic imputation procedure. We consider a variety of donor sampling schemes in the next three examples, for the case without covariates. More practically useful applications involving observed covariates are considered in Examples 4.9–4.12.

Example 4.8 *The Hot Deck by Simple Random Sampling with Replacement.* Let \bar{y}_{HD1} denote the hot deck estimator (4.3) when $\{H_i\}$ are obtained by random sampling with replacement from the observed values of Y . Conditioning on the sampled and observed values, the distribution of $\{H_1, \dots, H_r\}$ in repetitions of the hot deck is multinomial with sample size $n - r$ and probabilities $(1/r, \dots, 1/r)$ (see Cochran 1977, section 2.8). Hence, the moments of the distribution of $\{H_1, \dots, H_r\}$ given the observed data $Y_{(0)}$ are

$$E(H_i | Y_{(0)}) = (n - r)/r,$$

$$\text{Var}(H_i | Y_{(0)}) = (n - r)(1 - 1/r)/r,$$

$$\text{Cov}(H_i, H_j | Y_{(0)}) = -(n - r)/r^2.$$

Hence, taking expectations of the distribution of \bar{y}_{HD} over the distribution of $\{H_1, \dots, H_r\}$ yields:

$$E(\bar{y}_{\text{HD1}} | Y_{(0)}) = \bar{y}_R, \tag{4.6}$$

the respondent sample mean of Y , and

$$\text{Var}(\bar{y}_{\text{HD1}} | Y_{(0)}) = (1 - r^{-1})(1 - r/n)s_R^2/n, \tag{4.7}$$

where s_R^2 is the respondent sample variance of Y . In particular, assuming simple random sampling from a finite population of size N and missing data that are MCAR, (4.4) and (4.5) yield

$$E(\bar{y}_{\text{HD1}}) = \bar{Y}, \quad \text{Var}(\bar{y}_{\text{HD1}}) = (r^{-1} - N^{-1})S^2 + (1 - r^{-1})(1 - r/n)S^2/n, \tag{4.8}$$

where \bar{Y} and S^2 are the population mean and variance of Y , the first component of the variance is the simple random sample variance of \bar{y}_R , and the second component represents the increase in variance from the hot-deck procedure. The advantage of the hot deck method is that, unlike mean imputation, the distribution of the sampled values of Y is not distorted by the imputations.

The added sampling variance (4.7) from sampling imputations with replacement is a nonnegligible quantity. Specifically, the proportionate increase in the sampling variance of \bar{y}_{HD1} over \bar{y}_R is at most 0.25, and this maximum is attained when $r/n = 0.5$. Reductions in the additional variance from hot-deck

imputation can be achieved by a more efficient choice of sampling scheme, such as sampling *without replacement* (Problem 4.11), placing restrictions on the number of times a respondent acts as a donor, using the y -values themselves to form sampling strata (Bailar and Bailar 1983; Kalton and Kish 1981), systematic sampling from the ordered Y values, or using a sequential hot deck (Problem 4.11). However, we prefer multiple imputations, as discussed in Chapter 5, to these approaches because it not only can reduce the increase in sampling variance to negligible levels but also provides valid standard errors that take into account imputation uncertainty.

The hot deck estimators in Example 4.8 are unbiased only under the generally unrealistic MCAR assumption. If covariate information is available for responding and nonresponding units, then this information can be used to reduce nonresponse bias, as in the next two examples.

Example 4.9 Hot Deck Within Adjustment Cells. Suppose adjustment cells are formed and missing values within each cell are replaced by observed values from the same cell. Considerations relating to the choice of cells are similar to those in the choice of weighting classes for weighting estimates, discussed in Section 3.3.2. The mean and variance of the resulting hot deck estimates of \bar{Y} can be found by applying previous formulas separately in each cell and then combining over cells. Because adjustment cells are formed from the joint levels of categorical variables, they are not ideal for interval-scaled variables.

For example, the U.S. Census Bureau used this method for imputing earnings variables in the Income Supplement of the Current Population Survey (CPS) (Hanson 1978). For each nonrespondent on one or more income variables, the CPS hot deck finds a matching respondent based on variables that are observed for both units; the missing variables for the nonrespondent are then replaced by the respondent's values. The set of observed covariates is extensive, including age, race, sex, family relationship, children, marital status, occupation, schooling, full/part time, type of residence, and income recipiency pattern, so their joint classification creates a large matrix. When no match can be found for a nonrespondent, the CPS hot deck searches for a match at a lower level of detail, obtained by omitting some variables and collapsing the categories of others. David et al. (1986) compared imputations from the CPS hot deck with imputations using a more parsimonious regression model for income.

Example 4.10 Hot Deck Based on a Matching Metric. A more general approach to hot-deck imputation is to define a metric $d(i, j)$ to measure distance between units, based on the values of observed variables $x_i = (x_{i1}, \dots, x_{iK})^T$, and then to choose imputed values that come from responding units close to the unit with the missing value. That is, we choose an imputed value for y_i from a donor pool of units j that are such that (i) $y_j, x_{j1}, \dots, x_{jK}$ are observed and

(ii) $d(i, j)$ is less than some value d_0 . Varying the value of d_0 can control the number of available candidates j . A substantial literature on matching metrics exists in the context of observational studies where treated units are matched to control units (Rubin 1973a,b; Cochran and Rubin 1973; Rubin and Thomas 1992, 2000). The metric

$$d(i, j) = \begin{cases} 0, & i, j \text{ in same cell} \\ 1, & i, j \text{ in different cells} \end{cases}$$

yields the method of Example 4.9. For continuous x_i , another possible choice is the Mahalanobis distance: $d(i, j) = x_i^T S_{xx}^{-1} x_j$, where S_{xx} is an estimate of the covariance matrix of x_i . A better choice is *predictive mean matching*, which is based on the metric:

$$d(i, j) = (\hat{y}(x_i) - \hat{y}(x_j))^2, \quad (4.9)$$

where $\hat{y}(x_i)$ is the predicted value of Y from the regression of Y on the x 's computed using the complete units. This metric is superior to others because it weights predictors according to their ability to predict the missing variable. It is also available as an option in some multiple imputation software packages.

Simulation studies suggest that predictive mean matching provides some protection against misspecification of the regression of Y on x , but imputation based on parametric models is superior when good matches to donor units are not available, as when the sample size is small.

Example 4.11 Hot Decks for Multivariate Missing Data. Let $X = (X_1, \dots, X_q)$ denote the fully observed variables, including design variables, and let $Y = (Y_1, \dots, Y_p)$ denote the variables with possibly missing values. Suppose the components of Y are missing for the same set of units, so the data have just two missing-data patterns, complete and incomplete units. One possibility is to develop distinct univariate hot decks for each variable, with different donor pools and donors for each variable. This approach has the advantage that the donor pools can be tailored for each variable, for example by estimating a different predictive mean for each variable and creating the donor pools for each incomplete variable using the predictive mean matching metric. However, a consequence of this method is that associations between the imputed variables are not preserved.

An alternative method, which Marker et al. (2002) call the “single-partition, common-donor” hot deck is to create a single donor pool for each nonrespondent, using for example the multivariate analog of the predictive mean metric Eq. (4.9):

$$d(i, j) = (\hat{Y}(x_i) - \hat{Y}(x_j))^T \hat{\text{var}}(y \cdot x_i)^{-1} (\hat{Y}(x_i) - \hat{Y}(x_j)), \quad (4.10)$$

where $\hat{\text{var}}(y \cdot x_i)$ is the estimated residual covariance matrix of Y_i given x_i . A donor unit from this pool is used to impute all the missing variables for a non-respondent. This approach preserves associations within the set. Because the same metric is used for all the variables, the donors are not tailored to each variable.

Another approach that preserves associations between p variables, which we term the p -partition hot deck, is to create the donor pool for Y_j using adjustment cells (or more generally, a metric) that conditions on X and (Y_1, \dots, Y_{j-1}) , for $j = 2, \dots, p$, using the nonrespondent's previously imputed values of (Y_1, \dots, Y_{j-1}) , when matching donors to nonrespondent (Marker et al. 2002). This approach allows the metric to be tailored for each variable, and the conditioning on previously imputed variables in the metric provides some preservation of associations, although the degree of success depends on whether the metrics for each variable Y_j capture associations with X and (Y_1, \dots, Y_{j-1}) and the extent to which "close" matches can be found. For extensions of these ideas to more general patterns of missing data, see Marker et al. (2002) and Andridge and Little (2010).

Example 4.12 *Imputation Methods for Repeated Measures with Dropouts.* Longitudinal data are often subject to attrition when units leave the study prematurely. Let $y_i = (y_{i1}, \dots, y_{iK})$ be a $(K \times 1)$ complete data vector of outcomes for subject i , possibly incompletely observed. Write $y_i = (y_{(0),i}, y_{(1),i})$, where $y_{(0),i}$ = observed part of y_i , $y_{(1),i}$ = missing part of y_i . Define M_i such that $M_i = 0$ for complete units and $M_i = k$ if a subject drops out between the $(k-1)$ th and k th observation time, that is $y_{i1}, \dots, y_{i,k-1}$ are observed and y_{ik}, \dots, y_{iK} are missing.

The method known as "last observation carried forward" (LOCF) is sometimes applied in medical studies (see, for example, Pocock 1983). For unit i with $M_i = k$, missing values are imputed by the last recorded value for that unit, that is

$$\hat{y}_{it} = y_{i,k-1}, \quad t = k, \dots, K.$$

Although simple, this approach makes the often unrealistic assumption that the value of the outcome remains unchanged after dropout. Alternative imputation methods retain the advantage of simplicity but provide for unit and time effects. For example, Little and Su (1989) consider imputations for a panel survey of income based on a row plus column model fit. For unit i with $M_i = k$, let $\bar{y}_{\text{obsi},i} = (k-1)^{-1} \sum_{t=1}^{k-1} y_{it}$ denote the mean for the available measurements for subject i , and let $\bar{y}_{\text{obsi},+}^{(cc)} = r^{-1} \sum_{l=1}^r \bar{y}_{\text{obsi},l}$ be the corresponding mean averaged over the set of r complete units. Let $\bar{y}_{+t}^{(cc)} = r^{-1} \sum_{l=1}^r y_{lt}$ be the complete-case

mean for time point t . The prediction from a row plus column fit to the complete units is

$$\tilde{y}_{it} = \bar{y}_{\text{obsi},i} - \bar{y}_{\text{obsi},+}^{(cc)} + \bar{y}_{+t}^{(cc)},$$

where the column (time) mean $\bar{y}_{+t}^{(cc)}$ is modified by the row (unit) effect ($\bar{y}_{\text{obsi},i} - \bar{y}_{\text{obsi},+}^{(cc)}$). Adding a residual $y_{lt} - \tilde{y}_{lt}$ from a randomly drawn (or matched) unit l yields an imputed draw of the form

$$\hat{y}_{it} = \tilde{y}_{it} + (y_{lt} - \tilde{y}_{lt}), \quad \tilde{y}_{lt} = \bar{y}_{\text{obsi},l} - \bar{y}_{\text{obsi},+}^{(cc)} + \bar{y}_{+t}^{(cc)},$$

which simplifies to

$$\hat{y}_{it} = y_{lt} + (\bar{y}_{\text{obsi},i} - \bar{y}_{\text{obsi},l}), \quad t = k, \dots, K,$$

which is simply the simple hot-deck draw y_{lt} modified by a unit effect ($\bar{y}_{\text{obsi},i} - \bar{y}_{\text{obsi},l}$). A key assumption in this method is additivity of row and column effects. If additivity is more appropriately applied on a logarithmic scale, a multiplicative (row \times column) fit yields the alternative form:

$$\hat{y}_{it} = y_{lt} \times (\bar{y}_{\text{obsi},i} / \bar{y}_{\text{obsi},l}), \quad t = k, \dots, K.$$

4.4 Conclusion

Imputations should generally be

- (a) Conditional on observed variables, to reduce bias due to nonresponse, improve precision, and preserve association between missing and observed variables;
- (b) Multivariate, to preserve associations between missing variables;
- (c) Draws from the predictive distribution rather than means, to provide valid estimates of a wide range of estimands.

A key problem with all approaches discussed in this chapter is that inferences based on the filled-in data do not account for imputation uncertainty. Thus, standard errors computed from the filled-in data are systematically underestimated, p -values of tests are too significant, and confidence intervals are too narrow. In Chapter 5, we consider two approaches to this problem that have relatively general applicability – replication methods and multiple imputation.

Problems

- 4.1** Consider a bivariate sample with $n = 45$; $r = 20$ complete units, 15 units with only Y_1 recorded, and 10 units with only Y_2 recorded. The data are filled in using unconditional means, as in Section 4.2. Assuming MCAR, determine the percentage bias of estimates of the following quantities computed from the filled-in data: (a) the variance of Y_1 (σ_{11}); (b) the covariance of Y_1 and Y_2 (σ_{12}); and (c) the slope of the regression of Y_2 on Y_1 (σ_{12}/σ_{11}). You can ignore bias terms that are negligible in large samples.
- 4.2** Repeat Problem 4.1 when the missing values are filled in by Buck's (1960) method of Example 4.3 and compare the answers.
- 4.3** Describe the circumstances where Buck's (1960) method clearly dominates both complete-case analysis and available-case analysis.
- 4.4** Derive the expressions for the biases of Buck's (1960) estimators of σ_{jj} and σ_{jk} , stated in Example 4.3.
- 4.5** Suppose data are an incomplete random sample on Y_1 and Y_2 , where Y_1 given $\theta = (\mu_1, \sigma_{11}, \beta_{20-13}, \beta_{21-13}, \beta_{23-13}, \sigma_{22-13})$ is $N(\mu_1, \sigma_{11})$ and Y_2 given Y_1 and θ is $N(\beta_{20-13} + \beta_{21-13}Y_1 + \beta_{23-13}Y_1^2, \sigma_{22-13})$. The data are MCAR, the first r units are complete, the next r_1 units observe Y_1 only, and the last r_2 units observe Y_2 only. Consider the properties of Buck's method, applied to (a) Y_1 and Y_2 , and (b) Y_1 , Y_2 , and $Y_3 = Y_1^2$ (so that Y_3 has the same pattern as Y_1 and is imputed from the regression of Y_3 on Y_1 , Y_2), for deriving estimates of (i) the unconditional means $E(Y_1 | \theta)$ and $E(Y_2 | \theta)$, and (ii) the conditional means $E(Y_1 | Y_2, \theta)$, $E(Y_1^2 | Y_2, \theta)$, and $E(Y_2 | Y_1, \theta)$.
- 4.6** Show that Buck's (1960) method yields consistent estimates of the means when the data are MCAR.
- 4.7** Buck's method (Example 4.3) might be applied to data with both continuous and categorical variables, by replacing the categorical variables by a set of dummy variables, numbering one less than the number of categories. Consider properties of this method when (a) the categorical variables are fully observed and (b) the categorical variables are subject to missing values (Little and Rubin 1987, section 3.4.3).
- 4.8** Derive the expressions for large-sample bias in Table 4.1.

- 4.9** Derive the expressions for the large-sample sampling variance of the Cmean and Cdraw estimates of μ_2 in the discussion of Example 4.6.
- 4.10** Derive expressions (4.6)–(4.8) for the simple hot deck where imputations are by simple random sampling with replacement. Assuming r/N is small and large samples, show that the proportionate variance increase of \bar{y}_{HD1} over \bar{y}_R is at most 0.25, and this maximum is attained when $r/n = 0.5$. How do these numbers change when the hot deck is applied within adjustment cells?
- 4.11** Consider a hot deck like that of Example 4.8, except that imputations are by random sampling of donors *without* replacement. To define the procedure when there are fewer donors than recipients, write $n - r = kr + t$, where k is a nonnegative integer and $0 < t < r$. The hot deck without replacement selects all the recorded units k times and then selects t additional units randomly without replacement to yield the $n - r$ values required for the missing data. Thus

$$\bar{y}_{\text{NR}}^* = (kr\bar{y}_R + t\bar{y}_t)/(n - r),$$

where \bar{y}_t is the mean of the t supplementary values of Y . If \bar{y}_{HD2} denotes the estimate of \bar{Y} from this procedure, then show that

$$E(\bar{y}_{\text{HD2}} | Y_{(0)}) = \bar{y}_R$$

and

$$\text{Var}(\bar{y}_{\text{HD2}} | Y_{(0)}) = (t/n)(1 - t/r)s_R^2/n.$$

Show that the proportionate variance increase of \bar{y}_{HD2} over \bar{y}_R is at most 0.125, and this maximum is attained when $k = 0$, $t = n/4$, and $r = 3n/4$.

- 4.12** Another method for generating imputations is the *sequential* hot deck, where responding and nonresponding units are treated in a sequence, and a missing value of Y is replaced by the nearest responding value preceding it in the sequence. For example, if $n = 6$, $r = 3$, y_1, y_4 , and y_5 are observed, and y_2, y_3 , and y_6 are missing, then y_2 and y_3 are replaced by y_1 , and y_6 is replaced by y_5 . If y_1 is missing, then some starting value is used (for example, a value chosen from records in a prior survey). This method formed the basis for early imputation schemes for the US Census Bureau's Current Population Survey.

Suppose a simple random sample and a Bernoulli missingness mechanism. Show that the sequential hot deck estimate of Y , say \bar{y}_{HD3} , is unbiased for \bar{Y} with sampling variance (for large r and n and ignoring finite population corrections) given by

$$\text{Var}(\bar{y}_{\text{HD3}}) = (1 + (n - r)/n)S^2/r.$$

Hence, the proportionate increase in sampling variance over \bar{y}_R is $(n - r)/n$, the fraction of missing data (see Bailer et al. 1978, for details).

- 4.13** Outline a situation where the “last observation carried forward” method of Example 4.12 gives poor estimates (see, for example, Little and Yau 1996).
- 4.14** For the artificial data sets generated for Problem 1.6, compute and compare estimates of the mean and variance of Y_2 from the following methods:
- (a) Complete-case analysis;
 - (b) Buck’s method, imputing the conditional mean of Y_2 given Y_1 from the linear regression based on complete units;
 - (c) Stochastic regression imputation based on the normal model, where a random normal deviate $N(0, s_{22.1}^2)$ is added to each of the conditional means from (b);
 - (d) Hot-deck imputation, with adjustment cells computed by categorizing the complete units into quartiles based on the distribution of Y_1 . Suggest a situation where (d) might be a superior method to (c).

5

Accounting for Uncertainty from Missing Data

5.1 Introduction

Chapter 4 is concerned with the point estimation of population quantities in the presence of missingness. In this chapter, we discuss estimates of uncertainty for these point estimates that incorporate the additional variability due to missingness. The sampling variance estimates presented here all effectively assume that the method of adjustment for missingness has succeeded in essentially eliminating bias due to missingness. In many applications, the issue of bias is more crucial than that of increased sampling variance. In fact, it can be argued that providing a valid estimate of sampling variance is worse than providing no estimate if the point estimator has a large bias that dominates the mean squared error.

We distinguish four general approaches to accounting for uncertainty from missing data:

1. Apply explicit sampling variance formulas that allow for missingness. For example, in Example 4.1, we showed that the weighting class estimator (3.6) is obtained by substituting means within adjustment cells. Thus, if selection is by simple random sampling, and missing data are missing completely at random (MCAR) within adjustment cells, then the explicit formula (3.7) for mean squared error can be applied to estimate the precision, with the corresponding large-sample confidence interval $\bar{y}_{wc} \pm z_{1-\alpha/2}\{\hat{mse}(\bar{y}_{wc})\}^{1/2}$. Equation (4.8) gives the sampling variance of the hot-deck estimator by simple random sampling with replacement (rswr), and this formula can be modified to yield an estimated sampling variance of the hot-deck estimator applied within adjustment cells. There may be scope for further development in this area, although it is doubtful whether explicit estimators can be found for complicated sequential hot-deck methods such as the one described in Example 4.3, except under overly simplified assumptions. We do not discuss this option further here.

2. Modify the imputations so that valid standard errors can be computed from a single filled-in data set. This approach is examined in Section 5.2 and Examples 5.1 and 5.2. The simplicity of this approach is attractive, but it lacks generality, and the modifications required to the imputations may compromise the quality of the point estimates themselves.
3. Apply the imputation and analysis procedure repeatedly to resampled versions of the incomplete data (Rao and Shao 1992; Rao 1996; Fay 1996; Shao et al. 1998; Shao 2002; Lee et al. 2002). Uncertainty is estimated from variability of point estimates from a suitable set of samples drawn from the original sample. Two major variants of resampling, the bootstrap and the jackknife, are examined in Section 5.3. These methods are often easy to implement and have broad applicability, but they rely on large samples and are computationally intensive.
4. Create multiply-imputed data sets that allow the additional uncertainty from imputation to be assessed. Any single imputation method involving draws from the predictive distribution of the missing values, as discussed in Section 4.3, can be converted into a multiple imputation (MI) method by creating multiple data sets with different sets of draws imputed. This idea is more computationally intensive than 1 or 2 but provides approximately valid standard errors under broad classes of imputation procedures. With MI, complete-data estimates and standard errors from each imputed data set are combined with estimates of between-imputation uncertainty derived from variability in the estimates across the data sets. MI has broad applicability and is less computationally demanding than the resampling methods, because the multiple data sets are only used to determine the added uncertainty from the incomplete data. The method is particularly useful for data-base construction, where a single database is being created for multiple users. Section 5.4 contains an introduction to MI, and theoretical underpinnings of the method are examined in Chapter 10.

Section 5.5 concludes the chapter with some comments on the relative merits of resampling and MI.

5.2 Imputation Methods that Provide Valid Standard Errors from a Single Filled-in Data Set

For sample surveys involving multistage sampling, weighting, and stratification, the calculation of valid estimates of sampling variance is not a simple task even with complete response. As a result, approximate methods have been developed that can be applied to estimates that are functions of means and totals for broad classes of sample designs. The simplicity of these methods arises from restricting calculations to quantities calculated for collections of the sampled

units known as ultimate clusters (UCs), which are the largest sampling units that are independently sampled at random from the population. For example, the first stage of a design to sample households may involve the selection of census enumeration areas (EAs). The sample may include some self-representing EAs, which are included in the sample with probability 1, and some nonself-representing EAs, which are sampled from the population of EAs. The ultimate clusters then consist of the nonself-representing EAs and the sampling units that form the first stage of subsampling of self-representing EAs.

Estimates of sampling variance calculated from estimates for UCs are based on Lemma 5.1:

Lemma 5.1 *Let $\hat{\theta}_1, \dots, \hat{\theta}_k$ be K random variables that are (i) uncorrelated and (ii) have common mean μ . Let*

$$\bar{\theta} = \frac{1}{k} \sum_{j=1}^k \hat{\theta}_j \quad \text{and} \quad \hat{v}(\bar{\theta}) = \frac{1}{k(k-1)} \sum_{j=1}^k (\hat{\theta}_j - \bar{\theta})^2.$$

Then (i) $\bar{\theta}$ is an unbiased estimate of μ and (ii) $\hat{v}(\bar{\theta})$ is an unbiased estimate of the variance of $\bar{\theta}$.

Proof: $E(\bar{\theta}) = \sum_{j=1}^k E(\hat{\theta}_j)/k = \mu$, proving (i). To show (ii), note that

$$\sum_{j=1}^k (\hat{\theta}_j - \bar{\theta})^2 = \sum_{j=1}^k (\hat{\theta}_j - \mu)^2 - k(\bar{\theta} - \mu)^2.$$

Hence

$$\begin{aligned} E\left(\sum_{j=1}^k (\hat{\theta}_j - \bar{\theta})^2\right) - k(k-1)\text{Var}(\bar{\theta}) &= \sum_{j=1}^k \text{Var}(\hat{\theta}_j) - k \text{Var}(\bar{\theta}) - k(k-1)\text{Var}(\bar{\theta}) \\ &= \sum_{j=1}^k \text{Var}(\hat{\theta}_j) - k^2 \text{Var}(\bar{\theta}). \end{aligned} \tag{5.1}$$

But

$$k^2 \text{Var}(\bar{\theta}) = \text{Var}\left(\sum_{j=1}^k \hat{\theta}_j\right) = \sum_{j=1}^k \text{Var}(\hat{\theta}_j),$$

because the estimates $\{\hat{\theta}_j\}$ are uncorrelated. Hence, expression (5.1) equals zero, proving (ii). This lemma can be applied directly to linear estimators for sample designs that involve rswr of UCs, as in the next example. \square

Example 5.1 Standard Errors from Cluster Samples with Imputed Data. Suppose the population consists of K UCs, and the sample design includes k UCs by simple rswr. Let t_j denote the total for a variable Y in UC j , and suppose we estimate the population total for Y ,

$$T = \sum_{j=1}^K t_j,$$

by the Horvitz–Thompson estimate

$$\hat{t}_{\text{HT}} = \sum_{j=1}^k \hat{t}_j / \pi_j,$$

where the sum is over the UCs included in the sample (say $j = 1, \dots, k$), \hat{t}_j is an unbiased estimate of t_j , and π_j is the probability that UC j is selected. Then (i) \hat{t}_{HT} and $\{k\hat{t}_j / \pi_j, j = 1, \dots, k\}$ are all unbiased estimates of T , and (ii) the estimates $\{k\hat{t}_j / \pi_j, j = 1, \dots, k\}$ are uncorrelated because the method of random sampling is with replacement. Hence, by the lemma,

$$\hat{\nu}(\hat{t}_{\text{HT}}) = \sum_{j=1}^k \frac{(k\hat{t}_j / \pi_j - \hat{t})^2}{k(k-1)} \quad (5.2)$$

is an unbiased estimator of the variance of \hat{t} .

Suppose now we have missingness, and we derive estimates \hat{t}_j of the UC totals by one of the weighting or imputation techniques discussed in Chapters 3 or 4. We can still use (5.2) to estimate the sampling variance provided:

Condition 5.1 The estimates \hat{t}_j are unbiased for t_j , that is the imputation or weighting procedure leads to unbiased estimates within UC j ; and

Condition 5.2 The imputations or weighting adjustments are conducted *independently within each UC*.

Condition 5.2 is needed so that estimates \hat{t}_j remain uncorrelated, which is a key condition for applying the lemma. Thus, if imputation is carried out within adjustment cells, the cells must not include parts of different ultimate clusters. This principle may lead to unacceptably small samples within cells, particularly if the number of UCs is large. Thus, the requirement of a valid estimate of sampling variance based on this lemma conflicts with the need for a point estimator with acceptably small bias.

In practice, UCs are rarely sampled with replacement. If they are selected by simple random sampling without replacement (srswor), then UC estimates are negatively correlated, and estimates such as (5.2) based on the lemma overestimate the sampling variance, which typically leads to confidence intervals with greater than their nominal coverage – valid confidence intervals according to Neyman's (1934) definition. We might hope that multiplication by the finite population correction $(1 - k/K)$ would correct the overestimation, but in fact, this typically leads to underestimation. An unbiased estimate requires information from the second and higher stages of sampling. Thus, simple sampling variance estimates based on UCs require that the proportion of UCs sampled is small so that the overestimation introduced by the more efficient sampling without replacement can be ignored. This is often the case, or at least assumed, in practical sample designs.

Example 5.2 *Standard Errors from Stratified Cluster Samples with Imputed Data.* Most sampling designs also involve stratification in the selection of UCs. Again, assuming that the proportion of UCs sampled in each stratum is small, valid estimates of the sampling variance of linear statistics can be derived from UC estimates. Suppose that there are H strata, and let \hat{t}_{hj} be an unbiased estimate of the total t_{hj} for UC j in stratum h , for $h = 1, \dots, H, j = 1, \dots, K_h$. We can estimate t by

$$\hat{t} = \sum_{h=1}^H \sum_{j=1}^{k_h} \frac{\hat{t}_{hj}}{\pi_{hj}} = \sum_{h=1}^H \hat{t}_h, \quad (5.3)$$

where the summations are over the H strata and the k_h units sampled in stratum h , π_{hj} is the probability of selection of UC hj in stratum h , and \hat{t}_h is the estimate of the total for stratum h . The sampling variance of \hat{t} is estimated by

$$\hat{v}(\hat{t}) = \sum_{h=1}^H \sum_{j=1}^{k_h} \frac{(k_h \hat{t}_{hj}/\pi_{hj} - \hat{t}_h)^2}{k_h(k_h - 1)}. \quad (5.4)$$

In particular, with two UCs selected in each stratum, which is an especially popular design, the estimate of variance is

$$\hat{v}(\hat{t}) = \frac{1}{4} \sum_{h=1}^H (\hat{t}_{h1}/\pi_{h1} - \hat{t}_{h2}/\pi_{h2})^2.$$

Conditions for using these estimates with imputed data are the same as those for random sampling. That is, \hat{v} is unbiased if each \hat{t}_{hl} is unbiased for t_{hl} , and the imputations are carried out independently in each UC. We now consider alternative methods that relax the possibly severe restriction that the imputations across UC are independent.

5.3 Standard Errors for Imputed Data by Resampling

5.3.1 Bootstrap Standard Errors

A variety of methods compute standard errors from the variability of estimates based on repeated resampling of the observed data. We describe here the two most common variants of these methods, the bootstrap and the jackknife. There are theoretical relationships between the methods; indeed the jackknife can be derived from a Taylor series approximation of the bootstrap distribution of a statistic (Efron 1979).

Example 5.3 *The Simple Bootstrap for Complete Data.* Let $\hat{\theta}$ be a consistent estimate of a parameter θ based on a sample $S = \{i : i = 1, \dots, n\}$ of independent units. Let $S^{(b)}$ be a sample of size n obtained from the original sample S by simple rswr, and let $\hat{\theta}^{(b)}$ be the estimate of θ obtained by applying the original estimation method to $S^{(b)}$, where b indexes the drawn samples. Let $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ be the set of estimates obtained by repeating this procedure B times. The bootstrap estimate of θ is then the average of the bootstrap estimates:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}. \quad (5.5)$$

In situations where $\hat{\theta}$ has a substantial small-sample bias, a bootstrap estimate of the bias is $B^{-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}) = \hat{\theta}_{\text{boot}} - \hat{\theta}$. Subtracting this from $\hat{\theta}$ yields a bias-corrected estimate $\hat{\theta}_{\text{bc}} = 2\hat{\theta} - \hat{\theta}_{\text{boot}}$ that, under certain conditions, has bias of lower order than that of $\hat{\theta}$.

Our focus here is on using the bootstrap for estimating precision. Large sample precision can be estimated from the bootstrap distribution of the $\hat{\theta}^{(b)}$, which is estimated by the histogram formed by the bootstrap estimates $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$. In particular, the bootstrap estimate of the sampling variance of $\hat{\theta}$ is

$$\hat{V}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2. \quad (5.6)$$

It can be shown that under quite general conditions, \hat{V}_{boot} is a consistent estimate of the variance of $\hat{\theta}$ or $\hat{\theta}_{\text{boot}}$ as n and B tend to infinity. Thus, if the bootstrap distribution is approximately normal, a $100(1-\alpha)\%$ bootstrap large sample confidence interval for a scalar θ can be computed as

$$I_{\text{norm}}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{boot}}}, \quad (5.7)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the normal distribution. Alternatively, if the bootstrap distribution is nonnormal, a $100(1 - \alpha)\%$ bootstrap interval can be computed as

$$I_{\text{emp}}(\theta) = (\hat{\theta}^{(b,l)}, \hat{\theta}^{(b,u)}), \quad (5.8)$$

where $\hat{\theta}^{(b,l)}$ and $\hat{\theta}^{(b,u)}$ are the empirical $(\alpha/2)$ and $(1 - \alpha/2)$ quantiles of the bootstrap distribution of θ . Stable intervals based on (5.7) require bootstrap samples of the order of $B = 200$. Intervals based on (5.8) require much larger samples to be stable, for example $B = 2000$ or more (Efron 1994). Efron (1987, 1994) discusses refinements of (5.7) and (5.8) when the bootstrap distribution is not close to normal.

The bootstrap samples are readily generated as follows: let $m_i^{(b)}$ be the number of times that unit i is included in the b th bootstrap sample, with $\sum_{i=1}^n m_i^{(b)} = n$. Then for simple rswr,

$$\left(m_1^{(b)}, \dots, m_n^{(b)} \right) \sim \text{MNOM}(n; (n^{-1}, n^{-1}, \dots, n^{-1})), \quad (5.9)$$

a multinomial distribution with sample size n and n cells with equal probabilities $1/n$. Thus, $\theta^{(b)}$ can be computed by generating the counts (5.9) from a multinomial distribution and then applying the estimation procedure for $\hat{\theta}$ to the modified data, with unit i assigned a weight $m_i^{(b)}$. Some software packages automate this operation for common statistical procedures.

Example 5.4 *The Simple Bootstrap Applied to Data Completed by Imputation.* Suppose the data are a simple random sample $S = \{i: i = 1, \dots, n\}$ of units, but some units i are incomplete. Suppose that a consistent estimate $\hat{\theta}$ of a parameter θ is computed by filling in the missing values in $S^{(b)}$ using some imputation method Imp, yielding imputed data $\hat{S} = \text{Imp}(S)$, from which $\hat{\theta}$ is found from the filled-in data \hat{S} . Bootstrap estimates $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ can be computed as follows:

For $b = 1, \dots, B$:

- Generate a bootstrap sample $S^{(b)}$ from the original unimputed sample S , with weights as in (5.9).
- Fill in the missing data in $S^{(b)}$ by applying the imputation procedure, say Imp, to the bootstrap sample $S^{(b)}$, $\hat{S}^{(b)} = \text{Imp}(S^{(b)})$.
- Compute $\theta^{(b)}$ on the filled-in data $\hat{S}^{(b)}$ from (b).

Then, Eq. (5.6) provides a consistent estimate of the sampling variance of $\hat{\theta}$, and (5.7) or (5.8) can be used to generate confidence intervals for a scalar estimand. A key feature of this procedure is that the imputation procedure is applied B times, once to each bootstrap sample. Hence, the approach is computationally intensive. A simpler procedure is to apply the imputation procedure

just once to yield an imputed data set \hat{S} and then apply the bootstrap method to the filled-in data \hat{S} to estimate the variance of $\hat{\theta}$. However, this approach clearly does not propagate the uncertainty in the imputations and hence, does not provide a valid estimate of variance.

A second key feature is that the imputation method must yield a consistent estimate $\hat{\theta}$ for the true parameter. This is not required for Eq. (5.6) to yield a valid estimate of sampling error, but it is required for Eqs. (5.7) and (5.8) to yield appropriate confidence coverage and for tests to have the nominal size – see, in particular, Rubin's (1994) discussion of Efron (1994). For example, imputation by conditional draws, as discussed in Section 4.3, is needed to provide validity for a range of nonlinear estimands.

This approach assumes large samples. With moderate-sized data sets, it is possible that an imputation procedure that works for the full sample may need to be modified for one or more of the bootstrap samples. For example, if imputation is within adjustment cells and an adjustment cell for a particular bootstrap sample has nonrespondents but no respondents, then the adjustment cells must be pooled with similar cells, or some other modification applied.

5.3.2 Jackknife Standard Errors

Quenouille's jackknife method (see, for example, Miller 1974) historically predates the bootstrap and is widely used in survey sampling applications. It involves a particular form of resampling where estimates are based on dropping a single unit or set of units from the sample. As in the previous section, we present the basic form of the method for complete data and then discuss an application to incomplete data.

Example 5.5 *The Simple Jackknife for Complete Data.* Let $\hat{\theta}$ be a consistent estimate of an estimand θ based on a simple random sample $S = \{i: i = 1, \dots, n\}$ of units. Let $S^{(j)}$ be a subsample of size $n - 1$ obtained by dropping the j th unit from the original sample, and let $\hat{\theta}^{(j)}$ be the estimate of θ from this subsample. The quantity

$$\tilde{\theta}_j = n\hat{\theta} - (n - 1)\hat{\theta}^{(j)} \quad (5.10)$$

is called a pseudo-value. The jackknife estimate of θ is the average of the pseudo-values:

$$\hat{\theta}_{\text{jack}} = \frac{1}{n} \sum_{j=1}^n \tilde{\theta}_j = \hat{\theta} + (n - 1)(\hat{\theta} - \bar{\theta}), \quad (5.11)$$

where $\bar{\theta} = \sum_{j=1}^n \hat{\theta}^{(j)}/n$. The jackknife estimate of the sampling variance of either $\hat{\theta}$ or $\hat{\theta}_{\text{jack}}$ is

$$\hat{V}_{\text{jack}} = \frac{1}{n(n-1)} \sum_{j=1}^n (\tilde{\theta}_j - \hat{\theta}_{\text{jack}})^2 = \frac{n-1}{n} \sum_{j=1}^n (\hat{\theta}^{(j)} - \bar{\theta})^2. \quad (5.12)$$

Observe that the multiplier $(n-1)/n$ of $(\hat{\theta}^{(j)} - \bar{\theta})^2$ in Eq. (5.12) is larger than the corresponding multiplier $1/(B-1)$ of $(\hat{\theta}^{(b)} - \bar{\theta})^2$ in the bootstrap formula (5.6); this difference reflects the fact that the jackknife estimates of θ tend to be closer to $\hat{\theta}$ than the bootstrap estimates because they only differ from the computation of $\hat{\theta}$ by using one fewer unit. It can be shown that under certain conditions, (5.11) and (5.12) have properties analogous to those of the bootstrap. In particular, \hat{V}_{jack} is a consistent estimate of the sampling variance of either $\hat{\theta}$ or $\hat{\theta}_{\text{jack}}$ as n tends to infinity, and if $\hat{\theta}$ is consistent and the jackknife distribution is approximately normal, a $100(1-\alpha)\%$ confidence interval for a scalar θ can be computed as

$$I_{\text{norm}}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{jack}}} \quad (5.13)$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha)$ percentile of the normal distribution.

Example 5.6 *The Simple Jackknife Applied to Data Completed by Imputation.* Suppose, as in Example 5.4, the original data are a simple random sample $S = \{i: i = 1, \dots, n\}$, but some units i are incomplete. An estimator $\hat{\theta}$ of an estimand θ is computed by filling in the missing values in S using some imputation method Imp, yielding imputed data $\hat{S} = \text{Imp}(S)$, and then estimating θ using $\hat{\theta}$ applied to the filled-in data \hat{S} ; suppose the resulting estimator is consistent for θ . The jackknife can be implemented as follows:

For $j = 1, \dots, n$:

- (a) Delete unit j from S , yielding the sample $S^{(j)}$.
- (b) Fill in the missing data in $S^{(j)}$ by applying the imputation procedure Imp, yielding $\hat{S}^{(j)} = \text{Imp}(S^{(j)})$.
- (c) Compute $\hat{\theta}$ on the filled-in data $\hat{S}^{(j)}$ from (b); call its value $\hat{\theta}^{(j)}$.

Equations (5.10)–(5.12) then provide an estimate of the sampling variance of $\hat{\theta}$, and Eq. (5.13) generates a confidence interval for a scalar estimand that is valid, asymptotically. As with the bootstrap, a key feature is that imputations are recomputed on each jackknifed sample. To reduce the computations when n is large, the data can be divided into K blocks of J units, where $n = JK$, and then blocks of approximately K units dropped to form the jackknife estimates. The quantity K then replaces the quantity n in Eqs. (5.10)–(5.12).

Example 5.7 Standard Errors from Stratified Cluster Samples with Imputed Data (Example 5.2 Continued). The jackknife is quite commonly applied to sample surveys involving stratified multistage selection of units, as in the setting of Example 5.2 (Rao and Shao 1992; Rao 1996; Fay 1996). The jackknife where individual sample units are dropped does not yield valid standard errors because units within UCs tend to be correlated. Rather, the correct approach is to apply the jackknife with entire UCs deleted. Suppose interest is in a function $\theta = \theta(T)$ of a vector of population totals T . Suppose as in Example 5.2, there are H strata, and let \hat{t}_{hj} be an unbiased estimate of the total t_{hj} for UC j in stratum h , for $h = 1, \dots, H, j = 1, \dots, K_h$. With complete data S , we can estimate θ by $\hat{\theta} = \theta(\hat{t})$ where

$$\hat{t} = \hat{t}(S) = \sum_{h=1}^H \sum_{j=1}^{k_h} \frac{\hat{t}_{hj}(S)}{\pi_{hj}} = \sum_{h=1}^H \hat{t}_h(S),$$

say, where the summations are over the H strata and the k_h units are sampled in stratum h , π_{hj} is the probability of selection of UC hj in stratum h , and \hat{t}_h is the estimator of the vector of totals for stratum h . To apply the jackknife with no missingness, let $\hat{t}^{(\setminus h)}$ be the estimator of T with UC hj deleted. That is

$$\hat{t}^{(\setminus h)} = \sum_{h' \neq h}^H \hat{t}_{h'} + \hat{t}_h^{(\setminus h)}, \quad \text{where } \hat{t}_{h'} = \sum_{j'=1}^{k_{h'}} \frac{\hat{t}_{h'j'}}{\pi_{h'j'}}, \quad \hat{t}_h^{(\setminus h)} = \sum_{j' \neq j}^{k_h} \left(\frac{k_h}{k_h - 1} \right) \frac{\hat{t}_{hj'}}{\pi_{hj'}},$$

where contributions from UCs in stratum h other than hj have been multiplied by $k_h/(k_h - 1)$ to compensate for the dropped UC. The jackknife estimate of the sampling variance of $\hat{\theta} = \theta(\hat{t})$ is

$$\hat{V}_{\text{jack}} = \sum_{h=1}^H \frac{k_h - 1}{k_h} \sum_{j=1}^{k_h} (\hat{\theta}^{(\setminus h)} - \hat{\theta})^2, \quad (5.14)$$

where

$$\hat{\theta}^{(\setminus h)} = \theta(\hat{t}^{(\setminus h)}) \quad (5.15)$$

is the corresponding jackknifed estimate of θ . For linear estimators of scalar T and $\theta(T) = T$, the estimator (5.14) reduces to the previous estimator (5.4).

Now suppose there are missing values, and they are filled in by some imputation method Imp. The estimate of T becomes

$$\hat{t}(\hat{S}) = \sum_{h=1}^H \sum_{j=1}^{k_h} \frac{\hat{t}_{hj}(\hat{S})}{\pi_{hj}} = \sum_{h=1}^H \hat{t}_h(\hat{S}),$$

where $\hat{S} = \hat{S}(\text{Imp})$ denotes the imputed data set. As before, we assume that Imp is such that $\hat{t}(\hat{S})$ is a consistent estimator of T . The jackknife can then be implemented to estimate its sampling variance as follows:

For $h = 1, \dots, H$; $j = 1, \dots, k_h$:

- Delete UC hj from S , yielding the depleted sample $S^{(hj)}$.
- Fill in the missing data in $S^{(hj)}$ by applying the imputation procedure Imp to $S^{(hj)}$, yielding $\hat{S}^{(hj)} = I(S^{(hj)})$.
- Compute $\hat{t}^{(hj)}$ on the filled-in data $\hat{S}^{(hj)}$ from (b). That is

$$\hat{t}^{(hj)} = \sum_{h' \neq h}^H \hat{t}_{h'}(\hat{S}^{(hj)}) + \hat{t}_h^{(hj)}(\hat{S}^{(hj)}),$$

(5.16)

where $\hat{t}_{h'}(\hat{S}^{(hj)}) = \sum_{j'=1}^{k_{h'}} \frac{\hat{t}_{h'j'}(\hat{S}^{(hj)})}{\pi_{h'j'}}$, $\hat{t}_h^{(hj)} = \sum_{j' \neq j}^{k_h} \left(\frac{k_h}{k_h - 1} \right) \frac{\hat{t}_{hj'}(\hat{S}^{(hj)})}{\pi_{hj'}}$.

- Estimate the sampling variance of $\hat{t}(\hat{S})$ using Eqs. (5.14) and (5.15).

The cumbersome notation in (5.16) is used to emphasize the role of the changing nature of the imputed data sets for each jackknife sample. In particular, unlike the complete-data case, estimates of totals in strata other than h are potentially affected when UC hj is removed because imputations in those strata may be affected by the depletion of the donor pool. Rao and Shao (1992), Shao (2002), and Fay (1996) provide formulae for the adjustments to simple imputation methods that are needed when UCs are removed by the jackknife.

5.4 Introduction to Multiple Imputation

MI refers to the procedure of replacing each missing value by a vector of $D \geq 2$ imputed values. The D values are ordered in the sense that D completed data sets can be created from the vectors of imputations; replacing each missing value by the first component in its vector of imputations creates the first completed data set, replacing each missing value by the second component in its vector creates the second completed dataset, and so on. Standard complete data methods are used to analyze each completed data set. When the D sets of imputations are repeated random draws from the predictive distribution of the missing values under a particular model for missingness, the D completed data inferences can be combined to form one inference that properly reflects uncertainty due to missingness under that model. When the imputations are from two or more models for missingness, the set of inferences under each model can be contrasted across models to display the sensitivity

of inference to models for missingness, a particularly critical activity when MNAR missingness is being entertained.

MI was first proposed in Rubin (1978b), and a comprehensive treatment is given in Rubin (1987a), with additional appendix materials appearing in Rubin (2004). Other references include Herzog and Rubin (1983), Rubin (1986, 1996), Rubin and Schenker (1986), and more recent textbooks by van Buuren (2012) and Carpenter and Kenward (2014). The method has potential for application in a variety of contexts. It appears particularly promising in complex surveys with standard complete data analyses that are difficult to modify analytically in the presence of missingness. Here, we provide a brief overview of MI and illustrate its use.

As already indicated in Chapter 4, the practice of imputing for missing values is very common. Single imputation has the practical advantage of allowing standard complete data methods of analysis to be used. Imputation also has an advantage in many contexts in which the data collector (e.g., the Census Bureau) and the data analyst (e.g., a university social scientist) are different individuals, where the data collector may have access to more and better information about nonrespondents than the data analyst. For example, in some cases, information protected by confidentiality constraints (e.g., zip codes of dwelling units) may be available only to the data collector and can be used to help impute missing values (e.g., annual incomes). The obvious disadvantage of single imputation is that imputing a single value treats that value as known, and thus, without special adjustments, single imputation cannot reflect sampling variability under one model for missingness or uncertainty about the correct model for missingness.

MI shares both advantages of single imputation and rectifies both disadvantages. Specifically, when the D imputations are repetitions under one model for missingness, the resulting D complete data analyses can be easily combined to create an inference that validly reflects sampling variability because of the missing values under that model. When the MIs are from more than one model, uncertainty about the correct model is displayed by the variation in valid inferences across the models. The only disadvantage of MI over single imputation is that it takes more work to create the imputations and analyze the results, and more data storage. In today's computing environments, however, the extra storage requirements are often trivial, and the work in analyzing the data is quite modest because it basically involves performing the same task D times instead of once.

MIs ideally should be drawn according to the following protocol. For each model being considered, the D imputations of the missing values $Y_{(1)}$ are D repetitions from the posterior predictive distribution of $Y_{(1)}$, each repetition corresponding to an independent drawing of the parameters and missing values; "posterior predictive" meaning conditioning on the observed data $Y_{(0)}$ when predicting missing values. In practice, implicit models can often be used

in place of explicit models. Both types of models are illustrated in Herzog and Rubin (1983), where repeated imputations are created using (i) an explicit regression model and (ii) an implicit model, which is a modification of the Census Bureau's hot deck.

The analysis of a multiply-imputed data set is quite direct. First, each data set completed by imputation is analyzed using the same complete-data method that would be used in the absence of nonresponse. Let $\hat{\theta}_d, W_d, d = 1, \dots, D$ be D complete-data estimates and their associated sampling variances for a scalar estimand θ , calculated from D repeated imputations under one model. The combined estimate is

$$\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^D \hat{\theta}_d. \quad (5.17)$$

Because the imputations involved in MI are conditional draws rather than conditional means, under a good imputation model they provide valid estimates for a wide range of estimands, as discussed in Sections 4.3 and 4.4. Furthermore, the averaging over D imputed data sets in (5.17) increases the efficiency of estimation over that obtained from a single data set with conditional draws imputed.

The variability associated with estimate (5.17) has two components: the average within-imputation variance,

$$\overline{W}_D = \frac{1}{D} \sum_{d=1}^D W_d, \quad (5.18)$$

and the between-imputation component,

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta}_D)^2. \quad (5.19)$$

For vector θ , (5.17) and (5.18) are unchanged and $(\cdot)^2$ in (5.19) is replaced by $(\cdot)^T(\cdot)$. The total variability associated with $\bar{\theta}_D$ is

$$T_D = \overline{W}_D + \frac{D+1}{D} B_D, \quad (5.20)$$

where $(1 + 1/D)$ is an adjustment for finite due to estimating θ by $\bar{\theta}_D$. Hence,

$$\hat{\gamma}_D = (1 + 1/D)B_D/T_D \quad (5.21)$$

is an estimate of the fraction of information about θ missing due to nonresponse. For large sample sizes and scalar θ , the reference distribution for interval estimates and significance tests is a t distribution,

$$(\theta - \bar{\theta}_D)T_D^{-1/2} \sim t_v, \quad (5.22)$$

where the degrees of freedom,

$$\nu = (D - 1) \left(1 + \frac{1}{D+1} \frac{\bar{W}_D}{B_D} \right)^2, \quad (5.23)$$

is based on a Satterthwaite approximation (Rubin and Schenker 1986; Rubin 2004). An improved expression for the degrees of freedom for small data sets is

$$\nu^* = (\nu^{-1} + \hat{\nu}_{\text{obs}}^{-1})^{-1}, \quad \text{where } \hat{\nu}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \right) \nu_{\text{com}}, \quad (5.24)$$

and ν_{com} is the degrees of freedom for approximate or exact t inferences about θ when there are no missing values (Barnard and Rubin 1999). The theoretical basis for these expressions is examined further in Chapter 10.

For θ with K components, significance levels for null values of θ can be obtained from D repeated complete-data estimates $\hat{\theta}_d$ and variance–covariance matrices W_d using multivariate analogs of (5.17)–(5.20). Less precise p -values can be obtained directly from D repeated complete-data significance levels. Details are provided in Chapter 10.

Although MI is most directly motivated from the Bayesian perspective, the resultant inferences can be shown to possess good sampling properties. For example, Rubin and Schenker (1986) show that in many cases, interval estimates created using only two imputations provide randomization-based coverages close to their nominal levels even when there is as much as 30% missing information.

Example 5.8 *Multiple Imputation for Stratified Random Samples.* The main advantage of MI lies with more complex cases involving multivariate data with general patterns of missing data, but to illustrate the basics of the method, we consider inference for a population mean \bar{Y} from a stratified random sample. Suppose the population consists of H strata, and let N_h be the population size in stratum h , $N = \sum_{h=1}^H N_h$. Suppose a simple random sample of size n_h is taken in each stratum h , and let $n = \sum_{h=1}^H n_h$. With complete data, \bar{Y} is usually estimated by the stratified mean

$$\bar{y}_{\text{Strat}} = \sum_{h=1}^H P_h \bar{y}_h,$$

where \bar{y}_h is the sample mean and $P_h = N_h/N$ is the population proportion in stratum h . The estimated sampling variance of \bar{y}_{Strat} is

$$\text{Var}(\bar{y}_{\text{Strat}}) = \sum_{h=1}^H P_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_h^2}{n_h}, \quad (5.25)$$

where s_h^2 is the sample variance in stratum h .

Now suppose that only r_h of the n_h units in stratum h are respondents. With MI, each of the $\sum_{h=1}^H (n_h - r_h)$ nonresponding units would have D imputations, thereby creating D completed data sets and D values of the stratum means and variances, say, $\bar{y}_{h(d)}$ and $s_{h(d)}^2$, $d = 1, \dots, D$. The MI estimate (5.17) of \bar{Y} is the average of the D complete-data estimates of \bar{Y} :

$$\hat{\bar{Y}}_{\text{MI}} = \frac{1}{D} \sum_{d=1}^D \left(\sum_{h=1}^H P_h \bar{y}_{h(d)} \right). \quad (5.26)$$

From (5.18) to (5.20), the sampling variability associated with $\hat{\bar{Y}}_{\text{MI}}$ is the sum of the two components:

$$T_D = \frac{1}{D} \sum_{d=1}^D \sum_{h=1}^H P_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{s_{h(d)}^2}{n_h} + \frac{D+1}{D} \frac{1}{D-1} \sum_{d=1}^D \left(\sum_{h=1}^H P_h \bar{y}_{h(d)} - \hat{\bar{Y}}_{\text{MI}} \right)^2. \quad (5.27)$$

From (5.21) and (5.22), resulting inferences for \bar{Y} follow from the statement that $(\bar{Y} - \hat{\bar{Y}}_{\text{MI}})$ is distributed as t with center 0, squared scale given by (5.27), and degrees of freedom given by (5.23) with large samples and (5.24) otherwise.

If the missingness mechanism within each stratum is MCAR, imputation is not needed; the best estimate of \bar{Y} (in the absence of additional covariate information) is the stratified respondent mean

$$\hat{\bar{Y}}_{\text{Strat}} = \sum_{h=1}^H P_h \bar{y}_{hR}, \quad (5.28)$$

with associated sampling variance:

$$\text{Var}(\hat{\bar{Y}}_{\text{Strat}}) = \sum_{h=1}^H P_h^2 \left(1 - \frac{r_h}{N_h} \right) \frac{s_{hR}^2}{r_h}, \quad (5.29)$$

where \bar{y}_{hR} and s_{hR}^2 are the respondent mean and variance in stratum h . A desirable property of an MI method (in the absence of supplemental information) is that it reconstructs this estimate and associated variance as the D tends to infinity. Because the MIs are drawn from a predictive distribution, an intuitive method for creating imputations is the hot deck, which draws the nonrespondents' values at random from the respondents' values in the same stratum. Arguments in Rubin (1979) and Herzog and Rubin (1983) can be used to show that, for this method of imputation, $\hat{\bar{Y}}_{\text{MI}} \rightarrow \hat{\bar{Y}}_{\text{Strat}}$ as $D \rightarrow \infty$, so MI yields the

appropriate estimate as the number of imputations tends to infinity. However, the MI variance given by (5.27) is less than the sampling variance of the stratified respondent mean (5.29), even for infinite D . The source of the problem is that multiply-imputed draws using the hot deck do not reflect uncertainty about the stratum parameters. Simple modifications of the hot deck do reflect such uncertainty, and therefore with large D yield, not only the poststratified estimator but also the correct associated sampling variance.

First, consider a method based on an implicit model, which is called the *approximate Bayesian bootstrap* by Rubin and Schenker (1986). For $d = 1, \dots, D$, carry out the following steps independently: For each stratum, first create m_h possible values of Y by drawing m_h values at random with replacement from the r_h observed values of Y in stratum h ; and second, draw the m_h missing values of Y at random with replacement from these m_h values. Results in Rubin and Schenker (1986) or Rubin (2004) can be used to show that this method is “proper” for large D , in the sense that it will yield the stratified respondent mean (5.28) and its correct associated sampling variance (5.29) in this case. Another MI approach that yields the appropriate estimator (5.28) and associated sampling variance (5.29) is to impute using the Bayesian predictive distribution under a model in which Y -values within stratum h are assumed normal with mean μ_h and variance σ_h^2 , and $(\mu_h, \log \sigma_h)$ is assigned a uniform prior. The details of this procedure are deferred until Chapter 10.

5.5 Comparison of Resampling Methods and Multiple Imputation

The resampling methods of Section 5.3 and the MI method introduced in Section 5.4 are useful general tools for propagating imputation uncertainty. The relative merits of the approaches have been a subject of some debate. See for example the articles by Rubin (1996), Fay (1996), and Rao (1996) and associated discussion. We conclude this chapter with some general comments on this issue:

1. None of the methods is “model-free,” in the sense that they all make assumptions about the predictive distribution of the missing values in order to generate estimates based on the filled-in data that are consistent for population estimands (parameters).
2. In large samples where asymptotic arguments apply, resampling methods yield consistent estimates of sampling variance with minimal modeling assumptions, whereas MI estimates of sampling variance tend to be more closely tied to a particular model for the data and missingness mechanism. Thus, MI standard errors may be more appropriate for the particular data set when the model is sound, whereas resampling standard errors are

more generic (or less conditioned on features of the observed data) but are potentially less vulnerable to model misspecification. The issue of standard errors under misspecified models is discussed further in Chapter 6.

3. Resampling methods are based on large-sample theory, and their properties in small samples are questionable. The theory underlying MI is Bayesian and can provide useful inferences in small samples.
4. Some survey samplers have tended to be suspicious of MI because of its model-based Bayesian etiology and have favored sample-reuse methods because they derived under fewer parametric modeling assumptions. This issue may be more a question of complete-data analysis paradigms than of the method for dealing with imputation uncertainty. The relatively simple models of regression and ratio estimation with known covariates can form the basis for MI methods, and conversely, resampling standard errors can be computed for more complex parametric models. The right way to assess the relative merits of the methods, from a frequentist perspective, is through comparisons of their repeated-sampling operating characteristics in realistic settings, not their theoretical etiologies.
5. Some survey samplers have questioned the ability of MI to incorporate features of the sample design in propagating imputation uncertainty (e.g., Fay 1996). However, imputation models can incorporate stratification by including strata indicators as covariates and clustering by multilevel models that include random cluster effects. The complete-data inference can be design-based to incorporate these features, and can be based on a model that takes into account these features (Skinner et al. 1989).
6. MI is more useful than resampling methods for multiple-user database construction because a data set with a relatively small set of MIs (say 10 or fewer) can allow users to obtain excellent inferences for a broad range of estimands using complete-data methods, provided the MIs are based on a sound model (e.g., Ezzati-Rice et al. 1995). In contrast, resampling methods require 200 or more different imputed data sets, with imputations based on each resampled data set, and transmitting this large set of resampled and imputed data sets to users may not be practical. Thus in practice, the user needs software to implement a resampling imputation scheme on each replication.

Problems

- 5.1** As in Problem 1.6, generate 100 bivariate normal units $\{(y_{i1}, y_{i2}), i = 1, \dots, 100\}$ on (Y_1, Y_2) as follows:

$$\begin{aligned} y_{i1} &= 1 + z_{i1}, \\ y_{i2} &= 5 + 2^*z_{i1} + z_{i2}, \end{aligned}$$

where $\{(z_{i1}, z_{i2}), i = 1, \dots, 100\}$ are independent standard normal (mean 0, variance 1) deviates. The units (y_{i1}, y_{i2}) then form a bivariate normal sample with means (1, 5), variances (1, 5), and correlation $2/\sqrt{5} = 0.89$. Compute and compare estimated standard errors of estimates of (a) the mean of Y_2 and (b) the coefficient of variation of Y_2 , computed using the bootstrap, the jackknife and analytical formulae (exact for (a) or based on a large-sample approximation for (b)).

- 5.2** Create missing values of Y_2 for the data in Problem 5.1 by generating a latent variable U with values $u_i = 2^*(y_{i1} - 1) + z_{i3}$, where z_{i3} is a standard normal deviate, and setting y_{i2} as missing when $u_i < 0$. This mechanism is missing at random (MAR) because U depends on Y_1 but not Y_2 . With U having mean 0, about half of the values of Y_2 should be missing. Impute the missing values of Y_2 using conditional means from the linear regression of Y_2 on Y_1 , estimated from the complete units. Compute standard errors of estimates of the mean of Y_2 and coefficient of variation of Y_2 from the filled-in data, using the bootstrap and jackknife, applied both after imputation and before imputation; that is, for each replication of incomplete data, impute all missing values and estimate parameters. Which of these methods yield 90% intervals that actually cover the true parameter about 90% of the time? Which are theoretically valid, in the sense of yielding correct confidence interval coverage in large samples?
- 5.3** Repeat Problem 5.2 with the same observed data, but with missing values imputed using conditional draws rather than conditional means. That is, add a random normal deviate with mean 0 and variance given by the estimated residual variance, to the conditional mean imputations.
- 5.4** For the data in Problem 5.3, create ten multiply imputed data sets with different sets of conditional draws of the parameters, using the method of Problem 5.3. Compute 90% intervals for the mean and coefficient of variation of Y_2 using Eqs. (5.17)–(5.23), and compare with the single-imputation interval based on the first set of imputations. Estimate the fraction of missing information for each parameter estimate using Eq. (5.21).
- 5.5** As discussed in Section 5.4, the imputation method in Problem 5.4 is improper because it does not propagate the uncertainty in the regression parameter estimates. One way of making it proper is to compute the regression parameters for each set of imputations using a

bootstrap sample of the complete cases. Repeat Problem 5.4 with this modification and compare the resulting multiple-imputation intervals. How do they compare with the corresponding intervals from Problem 5.4?

- 5.6** Repeat Problem 5.4 or 5.5 for $D = 2, 5, 10, 20$, and 50 multiple imputes and compare answers. For what value of D does the inference stabilize?
- 5.7** Repeat Problem 5.4 or 5.5 using the more refined degrees of freedom formula (5.24) for the multiple imputation inference, and compare the resulting 90% nominal intervals with the simpler intervals based on (5.23).
- 5.8** Apply the methods in Problems 5.1–5.5 to 500 replicate data sets generated as in Problem 5.2 and assess the bias of the estimates and the coverage of intervals. Interpret the results given your understanding of the properties of the various methods.
- 5.9** Consider a simple random sample of size n with r respondents and $m = n - r$ nonrespondents, and let \bar{y}_R and s_R^2 be the sample mean and variance of the respondents' data, and \bar{y}_{NR} and s_{NR}^2 be the sample mean and variance of the imputed data. Show that the mean and variance \bar{y}_* and s_*^2 of all the data can be written as

$$\bar{y}_* = (r\bar{y}_R + m\bar{y}_{NR})/n \text{ and } s_*^2 = ((r-1)s_R^2 + (m-1)s_{NR}^2 + rm(\bar{y}_R - \bar{y}_{NR})^2/n)/(n-1).$$

- 5.10** Suppose in Problem 5.9, imputations are randomly drawn with replacement from the r respondents' values.
- (a) Show that \bar{y}_* is unbiased for the population mean \bar{Y} .
- (b) Show that conditional on the observed data, the variance of \bar{y}_* is $ms_R^2(1-r^{-1})/n^2$ and that the expectation of s_*^2 is $s_R^2(1-r^{-1})(1+rn^{-1}(n-1)^{-1})$.
- (c) Show that conditional on the sample sizes n and r (and the population Y -values), the variance of \bar{y}_* is the variance of \bar{y}_R times $(1+(r-1)n^{-1}(1-r/n)(1-r/N)^{-1})$, and show that this is greater than the expectation of $U_* = s_*^2(n^{-1} - N^{-1})$.
- (d) Assume r and N/r are large, and show that interval estimates of \bar{Y} based on U_* as the estimated variance of \bar{y}_* are too short by a factor $(1+nr^{-1}-rn^{-1})^{1/2}$. Note that there are two reasons: $n > r$ and \bar{y}_* is

not as efficient as \bar{y}_R . Tabulate true coverages and true significance levels as functions of r/n and nominal level.

- 5.11** Suppose multiple imputations are created using the method of Problem 5.10 D times, and let $\bar{y}_*^{(d)}$ and $U_*^{(d)}$ be the values of \bar{y}_* and U_* for the d th imputed data set. Let $\bar{\bar{y}}_* = \sum_{d=1}^D \bar{y}_*^{(d)}/D$, and T_* be the multiple imputation estimate of variance of \bar{y}_* . That is,

$$T_* = \bar{U}_* + (1 + D^{-1})B_*, \text{ where } \bar{U}_* = \sum_{d=1}^D U_*^{(d)}/D,$$

$$B_* = \sum_{d=1}^D (\bar{y}_*^{(d)} - \bar{\bar{y}}_*)^2.$$

- (a) Show that, conditional on the data, the expected value of B_* equals the variance of \bar{y}_* .
- (b) Show that the variance of $\bar{\bar{y}}_*$ (conditional on n, r , and the population Y -values) is $D^{-1} \operatorname{Var}(\bar{Y}_*) + (1 - D^{-1}) \operatorname{Var}(\bar{y}_R)$, and conclude that $\bar{\bar{y}}_*$ is more efficient than the single-imputation estimate \bar{y}_R .
- (c) Tabulate values of the relative efficiency of $\bar{\bar{y}}_*$ to \bar{y}_R for different values of D , assuming large r and N/r .
- (d) Show that the variance of \bar{y}_* (conditional on n, r , and the population Y -values) is greater than the expectation of T_* by approximately $s_R^2(1 - r/n)^2/r$.
- (e) Assume r and N/r are large, and tabulate true coverages and significance levels of the multiple-imputation inference. Compare with the results in Problem 10.3, part (d).

- 5.12** (a) Modify the multiple-imputation approach of Problem 5.11 to give the correct inferences for large r and N/r . (*Hint:* For example, add $s_R r^{-1/2} z_d$ to the imputed values for the d th multiply-imputed data set, where z_d is a standard normal deviate.)
- (b) Justify the adjustment in (a) based on the sampling variability of $(\bar{y}_R - \bar{Y})$.
- 5.13** Is multiple imputation (MI) better than imputation of a mean from the conditional distribution of the missing value because
- (i) it yields more efficient estimates from the filled-in data?
 - (ii) it yields consistent estimates of quantities that are not linear in the data?
 - (iii) it allows valid inferences from the filled-in data, if the imputation model is correct?
 - (iv) it yields inferences that are robust against model misspecification?

5.14 Is MI better than single imputation of a draw from the predictive distribution of the missing values (SI) because

- (i) it yields more efficient estimates from the filled-in data?
- (ii) unlike SI, it yields consistent estimates of quantities that are not linear in the data?
- (iii) it allows valid inferences from the filled-in data, if the imputation model is correct?
- (iv) it yields inferences that are robust against model misspecification?

Part II

Likelihood-Based Approaches to the Analysis of Data with Missing Values

6

Theory of Inference Based on the Likelihood Function

6.1 Review of Likelihood-Based Estimation for Complete Data

6.1.1 Maximum Likelihood Estimation

Many methods of estimation for incomplete data can be based on the likelihood function under specific modeling assumptions. In this section, we review basic theory of inference based on the likelihood function and describe how it is implemented in the incomplete data setting. We begin by considering maximum likelihood and Bayes' estimation for complete data sets. Only basic results are given, and mathematical details are omitted. For more detailed material, see, for example Cox and Hinkley (1974) and Gelman et al. (2013).

Suppose that Y denotes the data, where Y may be scalar, vector-valued, or matrix-valued according to context. The data are assumed to be generated by a model described by a probability or density function $p_Y(Y = y | \theta) = f_Y(y | \theta)$, indexed by a scalar or vector parameter θ , where θ is known only to lie in a parameter space Ω_θ . The “natural” parameter space for θ is the set of values of θ for which $f_Y(y | \theta)$ is a proper density – for example, the whole real line for means, the positive real line for variances, or the interval from zero to one for probabilities. Unless stated otherwise, we assume the natural parameter space for θ . Given the model and parameter θ , $f_Y(y | \theta)$ is a function of Y that gives the probabilities or densities of various Y values.

Definition 6.1 The *likelihood function* $L_Y(\theta | y)$ is any function of $\theta \in \Omega_\theta$ proportional to $f_Y(y | \theta)$; for fixed observed y by definition, $L_Y(\theta | y) = 0$ for any $\theta \notin \Omega_\theta$.

Note that the likelihood function, or more briefly, the likelihood, is a function of the parameter θ for fixed data y , whereas the probability or density $f_Y(y | \theta)$ is a function of y for fixed θ . In both cases, the argument of the function is written

first. It is slightly inaccurate to speak of “the” likelihood function because it really consists of a set of functions that differ by any factor that does not depend on θ .

Definition 6.2 The *loglikelihood function* $\ell_Y(\theta | y)$ is the natural logarithm (\ln) of the likelihood function $L_Y(\theta | y)$.

It is somewhat more convenient to work with the loglikelihood than with the likelihood in many problems.

Example 6.1 *Univariate Normal Sample.* The joint density of n independent and identically distributed units, $y = (y_1, \dots, y_n)^T$, from a normal population with mean μ and variance σ^2 is

$$f_Y(y | \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2} \right\}.$$

The loglikelihood function is

$$\ell_Y(\mu, \sigma^2 | y) = \ln [f_Y(y | \mu, \sigma^2)],$$

or ignoring the additive constant,

$$\ell_Y(\mu, \sigma^2 | y) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2}, \quad (6.1)$$

considered as a function of $\theta = (\mu, \sigma^2)$ for fixed observed data y .

Example 6.2 *Exponential Sample.* The joint density of n independent and identically distributed scalar units from the exponential distribution with mean $\theta > 0$ is

$$f_Y(y | \theta) = \theta^{-n} \exp \left\{ -\sum_{i=1}^n \frac{y_i}{\theta} \right\}.$$

Hence, the loglikelihood of θ given data y is

$$\ell_Y(\theta | y) = \ln \left\{ \theta^{-n} \exp \left(-\sum_{i=1}^n \frac{y_i}{\theta} \right) \right\} = -n \ln \theta - \sum_{i=1}^n \frac{y_i}{\theta}, \quad (6.2)$$

considered as a function of θ for fixed observed data y .

Example 6.3 *Multinomial Sample.* Suppose $y = (y_1, \dots, y_n)^T$, where y_i is categorical and takes one of C possible values $c = 1, \dots, C$. Let n_c be the number of units for which $y_i = c$, with $\sum_{c=1}^C n_c = n$. Conditional on n , the counts

(n_1, \dots, n_C) have a multinomial distribution with index n and probabilities $\theta = (\pi_1, \dots, \pi_{C-1})$ and $\pi_C = 1 - \pi_1 - \dots - \pi_{C-1}$. Then,

$$f_Y(y | \theta) = \left(\frac{n!}{n_1! \cdots n_C!} \right) \left(\prod_{c=1}^{C-1} \pi_c^{n_c} \right) (1 - \pi_1 - \dots - \pi_{C-1})^{n_C}.$$

The loglikelihood of θ given observed counts $\{n_c\}$ is then

$$\ell_Y(\theta | y) = \left(\sum_{c=1}^{C-1} n_c \ln \pi_c \right) + n_C \ln(1 - \pi_1 - \dots - \pi_{C-1}). \quad (6.3)$$

An important special case is the binomial distribution, obtained when $C = 2$.

Example 6.4 Multivariate Normal Sample. Let $y = (y_{ij})$, where $i = 1, \dots, n$, $j = 1, \dots, K$, be a matrix representing an independent and identically distributed sample of n units from the multivariate normal distribution with mean vector $\mu = (\mu_1, \dots, \mu_K)$ and covariance matrix $\Sigma = \{\sigma_{jk}, j = 1, \dots, K; k = 1, \dots, K\}$. Thus, y_{ij} represents the value of the j th variable for the i th unit in the sample. The density of y is

$$f_Y(y | \mu, \Sigma) = (2\pi)^{-nK/2} |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (y_i - \mu) \Sigma^{-1} (y_i - \mu)^T \right\}, \quad (6.4)$$

where $|\Sigma|$ denotes the determinant of Σ , the superscript “T” denotes the transpose of a matrix or vector, and y_i denotes the row vector of observed values for unit i . The likelihood of $\theta = (\mu, \Sigma)$ is (6.4) considered as a function of (μ, Σ) for fixed observed y . Here and elsewhere, for simplicity we write $\theta = (\mu, \Sigma)$, a minor abuse of notation because μ is a vector and Σ is a matrix.

The loglikelihood of $\theta = (\mu, \Sigma)$ is then

$$\ell_Y(\mu, \Sigma | y) = -(n/2) \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^n (y_i - \mu) \Sigma^{-1} (y_i - \mu)^T.$$

Finding the value of θ that is most likely to have generated the observed data y , by maximizing the likelihood function, is a basic tool for model-based inference about θ . Suppose that for fixed data y , two possible values of θ are being considered, θ' and θ'' . Suppose further that $L_Y(\theta' | y) = 2L_Y(\theta'' | y)$; then, it is reasonable to say that the observed outcome y is twice as likely under $\theta = \theta'$ as under $\theta = \theta''$. More generally, consider a value of θ , say $\hat{\theta}$, such that $L_Y(\hat{\theta} | y) \geq L_Y(\theta | y)$ for all other possible θ ; the observed outcome y is then at least as likely under $\hat{\theta}$ as under any other value of θ being considered. In some sense, such a value of θ is the one that is best supported by the data. This perspective leads to interest in the value of θ that maximizes the likelihood function.

Definition 6.3 A *maximum likelihood (ML) estimate* of θ is a value of $\theta \in \Omega_\theta$ that maximizes the likelihood $L_Y(\theta | y)$, or equivalently, the loglikelihood $\ell_Y(\theta | y)$.

This definition is phrased to allow the possibility of more than one ML estimate. For many standard important models and datasets, however, the ML estimate is unique. If the likelihood function is differentiable and bounded above, typically the ML estimate can be found by differentiating the likelihood (or the loglikelihood) with respect to θ , setting the result equal to zero, and solving for θ . The resulting equation

$$D_\ell(\theta) = 0, \quad \text{where } D_\ell(\theta) = \partial \ell_Y(\theta | y) / \partial \theta$$

is called the *likelihood equation*, and the derivative of the loglikelihood, $D_\ell(\theta)$, is called the *score function*. Letting d be the number of components in θ , the likelihood equation is a set of d simultaneous equations, defined by differentiating $\ell_Y(\theta | y)$ with respect to all d components of θ .

Example 6.5 Exponential Sample (Example 6.2 Continued). The loglikelihood for a sample from the exponential distribution is given by (6.2). Differentiating with respect to θ gives the likelihood equation

$$-\frac{n}{\theta} + \sum_{i=1}^n \frac{y_i}{\theta^2} = 0.$$

Solving for θ gives the ML estimate $\hat{\theta} = \bar{y} = \sum y_i / n$, the sample mean.

Example 6.6 Multinomial Sample (Example 6.3 Continued). The loglikelihood for the multinomial sample is given by (6.3). Differentiating with respect to π_c gives the likelihood equation

$$\frac{\partial \ell_Y(\theta | y)}{\partial \pi_c} = \frac{n_c}{\pi_c} - \frac{n_C}{1 - \pi_1 - \dots - \pi_{C-1}} = 0,$$

from which it is clear that the ML estimate $\hat{\pi}_c \propto n_c$ for all c . Hence, $\hat{\pi}_c = n_c / n$, the sample proportion in category c .

Example 6.7 Univariate Normal Sample (Example 6.1 Continued). From (6.1), the loglikelihood for a normal sample of n units is

$$\begin{aligned} \ell_Y(\mu, \Sigma | y) &= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \\ &= -\frac{n}{2} \ln \sigma^2 - \frac{n(\bar{y} - \mu)^2}{2\sigma^2} - \frac{(n-1)s^2}{2\sigma^2}, \end{aligned}$$

where $\bar{y} = \sum_{i=1}^n y_i/n$ and $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n - 1)$, the sample variance. Differentiating with respect to μ and setting the result equal to zero at $\mu = \hat{\mu}$ and at $\sigma^2 = \hat{\sigma}^2$ gives $(\bar{y} - \hat{\mu})^2/\hat{\sigma}^2 = 0$, which implies that $\hat{\mu} = \bar{y}$. Differentiating with respect to σ^2 and setting the result equal to zero at $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$ gives

$$-\frac{n}{2\hat{\sigma}^2} + \frac{n(\bar{y} - \hat{\mu})^2}{2\hat{\sigma}^4} + \frac{(n-1)s^2}{2\hat{\sigma}^4} = 0,$$

which, because $\hat{\mu} = \bar{y}$, implies that $\hat{\sigma}^2 = (n-1)s^2/n$, the sample variance with divisor n rather than $n-1$, that is, “uncorrected” for the loss of a degree of freedom used to estimate the mean. Thus, we obtain the ML estimates

$$\hat{\mu} = \bar{y}, \quad \hat{\sigma}^2 = (n-1)s^2/n.$$

Example 6.8 *Multivariate Normal Sample (Example 6.4 Continued).* Standard calculations in multivariate analysis (cf. Wilks 1963; Rao 1972; Anderson 1965; Gelman et al. 2013) show that maximizing (6.4) with respect to μ and Σ yields

$$\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = (n-1)S/n,$$

where $\bar{y} = (\bar{y}_1, \dots, \bar{y}_K)$ is the row vector of sample means, and $S = (s_{jk})$ is the $(K \times K)$ sample covariance matrix with (j, k) th element $s_{jk} = \sum_{i=1}^n (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)/(n-1)$.

The following property of maximum likelihood is useful in many problems:

Property 6.1 Let $g(\theta)$ be a function of the parameter θ . If $\hat{\theta}$ is an ML estimate of θ , then $g(\hat{\theta})$ is an ML estimate of $g(\theta)$.

If $g(\hat{\theta})$ is a one-to-one function of θ , Property 6.1 follows trivially from noting that the likelihood function of $\phi = g(\theta)$ is $L_Y(g^{-1}(\phi) | y)$, which is maximized when $\phi = g(\hat{\theta})$. If $g(\theta)$ is not a one-to-one function of θ (for example, the first component of θ), the property follows by defining a new one-to-one function of θ , say $g^*(\theta) = (g(\theta), h(\theta))$, and applying the aforementioned argument to g^* .

Example 6.9 *A Conditional Distribution Derived from a Bivariate Normal Sample.* The data consist of n independent and identically distributed units

$(y_{i1}, y_{i2}), i = 1, \dots, n$, from the bivariate normal distribution with mean (μ_1, μ_2) and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

As in Example 6.8, the ML estimates are

$$\hat{\mu}_j = \bar{y}_j, \quad j = 1, 2,$$

$$\hat{\sigma}_{jk} = (n - 1)s_{jk}/n, \quad j, k = 1, 2,$$

where \bar{y}_1 and \bar{y}_2 are the sample means and $S = (s_{jk})$ is the sample covariance matrix. By properties of the bivariate normal distribution (e.g., Stuart and Ord 1994, section 7.22), the conditional distribution of y_{i2} given y_{i1} is normal with mean $\mu_2 + \beta_{21.1}(y_{i1} - \mu_1)$ and variance $\sigma_{22.1}$, where

$$\beta_{21.1} = \sigma_{12}/\sigma_{11} \text{ and } \sigma_{22.1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$$

are, respectively, the slope and residual variance from the regression of Y_2 on Y_1 . By Property 6.1, the ML estimates of these quantities are

$$\hat{\beta}_{21.1} = \hat{\sigma}_{12}/\hat{\sigma}_{11} = s_{12}/s_{11},$$

the least squares estimate of the slope, and

$$\hat{\sigma}_{22.1} = \hat{\sigma}_{22} - \hat{\sigma}_{12}^2/\hat{\sigma}_{11} = \text{RSS}/n,$$

where $\text{RSS} = \sum_{i=1}^n \{y_{i2} - \bar{y}_2 - \hat{\beta}_{21.1}(y_{i1} - \bar{y}_1)\}^2$ is the residual sum of squares from the regression based on the n sampled units.

The ML estimates of $\beta_{21.1}$ and $\sigma_{22.1}$ can also be derived directly from the likelihood based on the conditional distribution of Y_2 given Y_1 . The connection between ML estimation for the normal linear regression model and least squares applies more generally, as discussed in the next example.

Example 6.10 *Multiple Linear Regression, Unweighted, and Weighted.* The data consist of n units $(y_i, x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$ on an outcome variable Y and p predictor variables $X = (X_1, \dots, X_p)$. Assume that, given the n values of $x_i = (x_{i1}, \dots, x_{ip})$, the values of y_i are independent normal random variables with mean $\beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ and common variance σ^2 . The loglikelihood of $\theta = (\beta, \sigma^2)$, where $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, given observed data $\{(x_i, y_i), i = 1, \dots, n\}$ is

$$\ell_Y(\theta | y) = -(n/2)\ln \sigma^2 - \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 / (2\sigma^2).$$

Maximizing this expression with respect to θ , the ML estimates of $(\beta_0, \dots, \beta_p)$ are found to be the least squares estimates of the intercept and the regression coefficients. Specifically, let

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

be the $n \times (p + 1)$ matrix of predictors including the constant term and the vector of outcomes, respectively. Then,

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \quad (6.5)$$

The ML estimate of σ^2 is

$$\hat{\sigma}^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta})/n \equiv \text{RSS}/n, \quad (6.6)$$

where RSS is the residual sum of squares from the least squares regression, the generalization of RSS in Example 6.9. Because the divisor here is n rather than $n - p - 1$, the ML estimate of σ^2 again does not correct for the loss of degrees of freedom when estimating the $p + 1$ location parameters, $(\beta_0, \beta_1, \dots, \beta_p)$.

A simple but important extension is weighted multiple linear regression. Suppose the mean of y_i is still $\mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$, but now its variance is σ^2/w_i , where $\{w_i\}$ are known positive constants. Thus, $(y_i - \mu_i)\sqrt{w_i}$ are independent, identically distributed (iid) $N(0, \sigma^2)$, and the log-likelihood is

$$\ell_Y(\theta | y) = -(n/2)\ln \sigma^2 - \sum_{i=1}^n w_i(y_i - \mu_i)^2/(2\sigma^2).$$

Maximizing this function yields ML estimates given by the weighted least squares estimates

$$\hat{\beta} = (X^T W X)^{-1} (X^T W Y), \quad (6.7)$$

and

$$\hat{\sigma}^2 = (Y - X\hat{\beta})^T W (Y - X\hat{\beta})/n, \quad (6.8)$$

where $W = \text{Diag}(w_1, \dots, w_n)$.

Example 6.11 Generalized Linear Models. Suppose that the data again consist of n units $(y_i, x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$ of an outcome variable Y and p predictor variables X_1, \dots, X_p . A more general class of models assumes that, given

$x_i = (x_{i1}, \dots, x_{ip})$, the n values of y_i are an independent sample from a regular exponential family distribution:

$$f_Y(y_i | x_i, \beta, \phi) = \exp[(y_i \delta(x_i, \beta) - b(\delta(x_i, \beta)))/\phi + c(y_i, \phi)], \quad (6.9)$$

where $\delta(\cdot, \cdot)$ and $b(\cdot)$ are known functions that determine the distribution of y_i , and $c(y_i, \phi)$ is a known function indexed by a scale parameter ϕ . The mean of y_i is assumed related to the covariates x_i by the expression:

$$E(y_i | x_i, \beta, \phi) = g^{-1} \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right), \quad (6.10)$$

or

$$g(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \quad (6.11)$$

where $\mu_i = E(y_i | x_i, \beta, \phi)$, and $g(\cdot)$ is a known one-to-one function. The function $g(\cdot)$ is called the *link function* because it “links” the expectation of y_i , μ_i , to a linear combination of the covariates. The *canonical* link g_c is that for which

$$g_c(\mu_i) = \delta(x_i, \beta) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}, \quad (6.12)$$

and is obtained by setting $g(\cdot)$ equal to the inverse of the derivative of $b(\cdot)$ with respect to its argument (see Problems 6.6 and 6.7 for more details). This choice of link function is a natural starting point for modeling many data sets. Important specific models with their canonical links include the following:

Normal linear regression: $g_c = \text{identity}$, $b(\delta) = \delta^2/2$, $\phi = \sigma^2$;

Poisson regression: $g_c = \log$, $b(\delta) = \exp(\delta)$, $\phi = 1$; and

Logistic regression: $g_c = \text{logit}$, where $\text{logit}(\mu_i) = \log(\mu_i/(1 - \mu_i))$, $b(\delta) = \log(1 + \exp(\delta))$, $\phi = 1$.

The loglikelihood of $\theta = (\beta, \phi)$ given observed data (y_i, x_i) , $i = 1, \dots, n$ based on Eq. (6.9) is

$$\ell_Y(\theta | y) = \sum_{i=1}^n [(y_i \delta(x_i, \beta) - b(\delta(x_i, \beta)))/\phi + c(y_i, \phi)],$$

which for nonnormal cases does not generally have an explicit maximum. Numerical maximization can be achieved using an iterative algorithm, such as the Fisher scoring algorithm (McCullagh and Nelder 1989, section 2.5; Firth 1991, section 3.4).

Example 6.12 Normal Repeated Measures Models. In longitudinal studies, units are observed at different times and/or under different experimental conditions. The following general repeated measures model is given in Jennrich and Schluchter (1986) and builds on earlier work by Hartley and Rao (1967), Harville (1977), Laird and Ware (1982), and Ware (1985). Suppose that the complete data for unit i consist of K measurements $y_i = (y_{i1}, \dots, y_{iK})$ of an outcome variable Y and that y_i are independent with distribution

$$y_i \sim N_K(X_i\beta, \Sigma(\psi)), \quad (6.13)$$

where X_i is a known $(K \times m)$ design matrix for unit i , β is a $(m \times 1)$ vector of unknown regression coefficients, and the elements of the covariance matrix Σ are known functions of q unknown parameters ψ . The model thus incorporates a mean structure, defined by β and the set of design matrices $\{X_i\}$, and a covariance structure defined by the form of the covariance matrix Σ . The observed data consist of the design matrices $\{X_i\}$ and outcome measurements $\{y_i, i = 1, \dots, n\}$.

A large number of situations can be modeled by combining different choices of mean and covariance structures. Common covariance structures include:

Independence: $\Sigma = \text{Diag}_K(\psi_1, \dots, \psi_K)$, a diagonal $(K \times K)$ matrix,

Compound symmetry: $\Sigma = \psi_1 U_K + \psi_2 I_K$, ψ_1 and ψ_2 scalar, $U_K = (K \times K)$ matrix of ones, $I_K = (K \times K)$ identity matrix,

Autoregressive, lag 1: $\Sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_1 \psi_2^{|j-k|}$, ψ_1, ψ_2 scalars,

Banded: $\Sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_r$, where $r = |j - k| + 1$, $r = 1, \dots, K$,

Factor analytic: $\Sigma = \Gamma \Gamma^T + \psi_0$, with $\Gamma = (K \times q)$ the matrix of “factor loadings,” $\psi_0 = (K \times K)$ the diagonal matrix of “specific variances,” and $\psi = (\Gamma, \psi_0)$.

Random effects: $\Sigma = Z \psi^* Z^T + \sigma^2 I_K$, where Z is a $(K \times q)$ known matrix, ψ^* is a $(q \times q)$ dispersion matrix, σ^2 is a scalar, I_K the $K \times K$ identity matrix, and $\psi = (\psi^*, \sigma^2)$.

Unstructured: $\Sigma = (\sigma_{jk})$, with ψ representing the $v = K(K+1)/2$ elements of this matrix.

The mean structure is also flexible. If $X_i = I_K$, the $(K \times K)$ identity matrix, then $\mu_i = \beta^T$ for all i . Between-subject and within-subject effects are readily modeled through other choices of X_i .

The loglikelihood for the model (6.13) is

$$\ell_Y(\beta, \psi) = -0.5n \log |\Sigma(\psi)| - 0.5 \sum_{i=1}^n (y_i - X_i\beta)^T \Sigma^{-1}(\psi) (y_i - X_i\beta), \quad (6.14)$$

which is linear in the observed quantities $\{y_i, y_i^T y_i, i = 1, \dots, n\}$. This likelihood does not yield explicit ML estimates, except in simple special cases. For the

unstructured covariance structure $\Sigma(\psi) = \Sigma$, there is no explicit ML estimate for (β, Σ) , but explicit ML estimates are available for β given Σ and for Σ given β . Hence, an iterative ML solution is obtained by the following alternating conditional modes algorithm: Given estimates $(\beta^{(t)}, \psi^{(t)})$ at iteration t , estimates at iteration $t + 1$ are

$$\beta^{(t+1)} = \sum_{i=1}^n \left(X_i^T (\Sigma^{(t)})^{-1} X_i \right)^{-1} X_i^T (\Sigma^{(t)})^{-1} y_i, \quad \text{and} \quad (6.15)$$

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \left(y_i - X_i^T \beta^{(t+1)} \right)^T \left(y_i - X_i^T \beta^{(t+1)} \right). \quad (6.16)$$

For other covariance structures, (6.15) remains unchanged, and (6.16) is replaced by an update for ψ that depends on the specific form of the structure.

A useful variant of ML is *restricted* ML, which yields estimates that correct for the loss of degrees of freedom from estimating location parameters in normal models, a correction that is not made with ML estimates, as noted in Examples 6.7, 6.8, and 6.10. Restricted ML was originally formulated for analysis of variance (ANOVA) models and linear mixed models (see, for example, Harville 1977) as the likelihood corresponding to the largest set of independent error contrasts of the data, where a contrast is a linear combination of the data with mean zero. More generally, it can be viewed as maximizing the function obtained by including a Bayesian prior distribution for the location parameters β and then integrating these parameters out of the product of the prior distribution and the likelihood function before maximizing the resulting function of ψ .

Other modifications of maximum likelihood, including conditional likelihood, marginal likelihood, or partial likelihood, maximize pieces of the likelihood rather than the full likelihood, generally to eliminate nuisance parameters and thereby simplify the problem. In the missing data context, this strategy can be applied to avoid modeling the missingness mechanism in certain situations, at the expense of some loss of information.

6.1.2 Inference Based on the Likelihood

By ML *inference*, we mean the ML estimate and a method for assessing its uncertainty, via hypothesis tests, confidence intervals, or Bayesian posterior distributions. We distinguish between (i) “pure” or “direct” likelihood inference (Edwards 1992; Royall 1997; Frumento et al. 2016), which involves the data solely through likelihood ratios $L_Y(y|\theta)/L_Y(y|\theta^*)$ for pairs of parameter values θ, θ^* , with data y fixed at its observed value and (ii) frequentist likelihood inference, meaning tests, or confidence intervals based on the repeated sampling distribution of the ML estimate.

The most popular form of direct likelihood inference is Bayesian inference, where the parameter θ is assigned a prior distribution $p(\theta)$, and inference about

θ , after observing the data y , is based on its posterior distribution $p(\theta | y)$, determined by Bayes' theorem:

$$p(\theta | y) = \frac{p(\theta)L_Y(y | \theta)}{p(y)}, \quad (6.17)$$

where $p(y) = \int p(\theta)L_Y(y | \theta)d\theta$ is the normalizing constant.

Point estimates of θ can be obtained as measures of the “center” of the posterior distribution, such as the posterior mean, median, or mode. In a Bayesian analysis, which always includes a prior distribution for θ in the model specification, we denote the mode of the posterior distribution $p(\theta | y)$ by $\hat{\theta}$. It corresponds to the ML estimate when the prior distribution is uniform:

$$p(\theta) = \text{const.} \quad \text{for all possible } \theta.$$

The latter is not a real probability distribution unless the parameter space has compact support, but it can be used to approximate the lack of prior knowledge about θ , providing care is taken to ensure that the resulting posterior distribution is well defined.

There are strong parallels between Bayesian and frequentist likelihood inference in the case of large samples, to which we now turn.

6.1.3 Large Sample Maximum Likelihood and Bayes Inference

In this section, we outline some basic large-sample properties of frequentist ML and Bayesian inference. References to these results include Huber (1967), DeGroot (1970), Rao (1972), Cox and Hinkley (1974), White (1982), and Gelman et al. (2013).

Let $\hat{\theta}$ denote an ML estimate of θ based on observed data y , or the posterior mode in a Bayesian analysis, and suppose that the model is correctly specified. The most important practical property of $\hat{\theta}$ is that, in many cases, especially with large samples, the following approximation is applicable.

Approximation 6.1

$$(\theta - \hat{\theta}) \sim N(0, C), \quad (6.18)$$

where C is an estimate of the $d \times d$ covariance matrix for $(\theta - \hat{\theta})$.¹

This approximation has both a frequentist and a Bayesian interpretation. The Bayesian version of Approximation 6.1 treats θ as the random variable and $\hat{\theta}$ as the mode of the posterior distribution, fixed by the observed data. The interpretation of (6.18) is then that given $f(\cdot | \cdot)$ and conditional on the observed values of the data, the posterior distribution of θ is normal with mean $\hat{\theta}$ and covariance

matrix C , where $\hat{\theta}$ and C are statistics fixed at their observed values. The theoretical justification is based on a Taylor series expansion of the loglikelihood about the ML estimate, namely,

$$\ell_Y(\theta | y) = \ell_Y(\hat{\theta} | y) + (\theta - \hat{\theta})^T D_{\ell}(\hat{\theta} | y) - \frac{1}{2}(\theta - \hat{\theta})^T I_Y(\hat{\theta} | y)(\theta - \hat{\theta}) + r(\theta | y),$$

where $D_{\ell}(\theta | y)$ is the score function, and $I_Y(\theta | y)$ is the observed information:

$$I_Y(\theta | y) = -\frac{\partial^2 \ell_Y(\theta | y)}{\partial \theta \partial \theta}.$$

By definition, $D_{\ell}(\hat{\theta} | y) = 0$. Hence, provided the remainder term $r(\theta | y)$ can be neglected, and the prior distribution of θ is relatively flat in the range of θ supported by the data, the posterior distribution of θ has density

$$f_Y(\theta | y) \propto \exp \left[-\frac{1}{2}(\theta - \hat{\theta})^T I_Y(\hat{\theta} | y)(\theta - \hat{\theta}) \right],$$

which is the normal distribution of Approximation 6.1 with covariance matrix

$$C = I_Y^{-1}(\hat{\theta} | y),$$

the inverse of the observed information evaluated at $\hat{\theta}$.

Technical regularity conditions for these results are discussed in the cited texts. Gelman et al. (2013) describe the following situations where the results can be expected to fail:

S1: “Underidentified” models, where the information on one or more parameters in the likelihood does not increase with the sample size.

S2: Models where the number of parameters increases with the sample size. We assume that the number of components of θ does not increase with the sample size, and hence do not consider “nonparametric” or “semiparametric” models where the number of parameters increases with the sample size. Asymptotic results for such models are tricky and require that the rate of increase of the number of parameters be carefully controlled.

S3: Unbounded likelihoods, where there is no ML estimate or posterior mode in the interior of the parameter space. This problem can often be avoided by not allowing isolated poles of the likelihood function, for example, by bounding variance parameters away from zero.

S4: Improper posterior distributions, which arise in some settings when improper prior distributions are specified. See, for example, Gelman et al. (2013).

S5: Points of convergence on the edge of the parameter space or with no support from the prior distribution. These problems can be avoided by checking the

plausibility of the model and giving positive prior probability to all values of the parameter, even those that are remotely plausible.

S6: Slow convergence in the tails of the distribution. The normal limiting distribution in Approximation 6.1 implies exponentially decaying tails, and this property may not be achieved quickly enough to provide valid inferences for realistic sample sizes. In such cases, a better approach may be to avoid the asymptotic approximation and concentrate on estimating the posterior distribution for a reasonable choice (or choices) of prior distributions. Alternatively, a transformation of the parameters may improve Approximation 6.1. This motivates the following.

Property 6.2 Let $g(\theta)$ be a monotone differentiable function of θ , and let C be the estimated large sample covariance matrix of $\theta - \hat{\theta}$, as in Approximation 6.1. Then, an estimate of the large-sample covariance matrix of $g(\theta) - g(\hat{\theta})$ is

$$D_g(\hat{\theta}) C D_g(\hat{\theta})^T,$$

where $D_g(\theta) = \partial g(\theta)/\partial\theta$ is the partial derivative of g with respect to θ . Property 6.2 follows from the first term of a Taylor series expansion of $g(\theta)$ about $\theta = \hat{\theta}$ and leads to the following approximation.

Approximation 6.2

$$g(\theta) - g(\hat{\theta}) \sim N[0, D_g(\hat{\theta}) C D_g(\hat{\theta})^T]. \quad (6.19)$$

When sample sizes are not large, it is often useful to find a transformation $g(\cdot)$ that makes the normality in Approximation 6.2 more accurate than in Approximation 6.1, for example, using $\ln(\sigma^2)$ instead of σ^2 in Example 6.7 or Fisher's normalizing transformation $Z_\rho = \ln((1+\rho)/(1-\rho))/2$ instead of the correlation $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ in Example 6.9.

The frequentist interpretation of Approximation 6.1 is that under $f(\cdot | \cdot)$ in repeated samples with fixed θ , $\hat{\theta}$ will be approximately normally distributed with mean equal to the true value of θ and covariance matrix C , which has lower-order variability than $\hat{\theta}$. The theoretical justification first approximates $D_\ell(\hat{\theta} | y)$ about the true value of θ by the first term of a Taylor series:

$$0 = D_\ell(\hat{\theta} | y) = D_\ell(\theta | y) - I_Y(\theta | y)(\hat{\theta} - \theta) + r(\hat{\theta} | y).$$

If the remainder term $r(\hat{\theta} | y)$ is negligible, we have

$$D_\ell(\theta | y) \approx I_Y(\theta | y)(\hat{\theta} - \theta).$$

Under certain regularity conditions, it can be shown by a central limit theorem that in repeated sampling, $D_\ell(\theta | y)$ is asymptotically normal with mean 0 and covariance matrix

$$J_Y(\theta) = E(I_Y(\theta | y) | \theta) = \int I_Y(\theta | y) f_Y(y | \theta) dy,$$

which is called the expected information matrix. A version of the law of large numbers implies that

$$J_Y(\theta) \cong J_Y(\hat{\theta}) \cong I_Y(\hat{\theta} | y).$$

Combining these facts leads to Approximation 6.1, with covariance matrix

$$C = J_Y^{-1}(\hat{\theta}),$$

the inverse of the expected information evaluated at $\theta = \hat{\theta}$, or

$$C = I_Y^{-1}(\hat{\theta} | y),$$

the inverse of the observed information evaluated at $\theta = \hat{\theta}$. Ancillarity arguments (Efron and Hinkley 1978) suggest that the observed information $I_Y^{-1}(\hat{\theta} | y)$ provides a better estimate of precision than the expected information $J_Y^{-1}(\hat{\theta})$. If a transformation of θ is applied to improve normality, these choices of C can also be substituted in Eq. (6.15) to obtain a frequentist version of Approximation 6.2.

Approximations 6.1 and 6.2 are based on the assumption that the model is correctly specified, that is, that Y is sampled from the density $f_Y(y | \theta_0)$ for some true value θ_0 of the parameter. If the model is misspecified, and Y is in fact sampled from the true density $f_Y^*(Y)$, the posterior mode or the ML estimate converges to the value θ^* of θ that maximizes the Kullback–Liebler Information $E[\log(f_Y(y | \theta)/f_Y^*(y))]$ of the model distribution $f_Y(y | \theta)$ with respect to the true distribution $f_Y^*(y)$. The frequentist version of Approximation 6.1 can then be replaced by:

Approximation 6.3

$$(\hat{\theta} | f^*) \sim N(\theta^*, C^*), \quad (6.20)$$

where θ^* is previously defined and

$$C^* = J_Y^{-1}(\theta) K_Y(\theta) J_Y^{-1}(\theta),$$

where $K_Y(\theta) = E(D_\ell(\theta)D_\ell(\theta)^T)$. When the model is correctly specified, θ^* equals the true value of θ , namely θ_0 , and C^* reduces to $J^{-1}(\theta_0)$ (White 1982). In situations where the model is incorrectly specified, but θ^* is nevertheless the parameter of interest, as when $\hat{\theta}$ is consistent for θ , the covariance matrix of $\hat{\theta}$ can then be consistently estimated by the so-called *sandwich* estimator of C^* :

$$\hat{C}^* = I_Y^{-1}(\hat{\theta} | y)\hat{K}_Y(\hat{\theta})I_Y^{-1}(\hat{\theta} | y), \quad \text{where } \hat{K}_Y(\hat{\theta}) = D_\ell(\hat{\theta})D_\ell(\hat{\theta})^T. \quad (6.21)$$

This estimator is less precise than the observed or expected information but provides some protection against model misspecification. See Gelman et al. (2013) for more details. For a Bayesian interpretation of the sandwich estimator, see Szapiro et al. (2010).

Another method for calculating variances takes bootstrap samples of the data and calculates the ML estimates on each. The sample variance of these bootstrap estimates is asymptotically equivalent to the variance computed using (6.21). The jackknife can also be used to provide asymptotic standard errors. An introduction to these methods is given in Section 5.3. For more discussion, see Efron and Tibshirani (1993) and Miller (1974).

From the Bayesian or frequentist perspectives, Approximations 6.1 or 6.2, with C fixed at $I_Y^{-1}(\hat{\theta})$, $J_Y^{-1}(\hat{\theta})$, the sandwich estimator (6.21), or some other approximation, can be used to provide interval estimates for θ . For example, 95% intervals for scalar θ are given by

$$\hat{\theta} \pm 1.96C^{1/2}, \quad (6.22)$$

where 1.96 can often be replaced by 2 in practice. For vector θ , 95% ellipsoids are given by the inequality

$$(\theta - \hat{\theta})^T C^{-1}(\theta - \hat{\theta}) \leq \chi_{0.95,d}^2, \quad (6.23)$$

where $\chi_{0.95,d}^2$ is the 95th percentile of the chi-squared distribution with degrees of freedom d , the dimensionality of θ . More generally, 95% confidence ellipsoids for $q < d$ components of θ , say $\theta_{(1)}$, can be calculated as

$$(\theta_{(1)} - \hat{\theta}_{(1)})^T C_{(11)}^{-1}(\theta_{(1)} - \hat{\theta}_{(1)}) \leq \chi_{0.95,q}^2, \quad (6.24)$$

where $\hat{\theta}_{(1)}$ is the ML estimate of $\theta_{(1)}$, and $C_{(11)}$ is the submatrix of C corresponding to $\theta_{(1)}$.

Under $f_Y(\cdot | \cdot)$ and assuming large enough samples to make Approximation 6.1 appropriate, inference based on Eq. (6.18) is not only appropriate but also asymptotically optimal. It is thus not surprising that the ML estimator for θ with Approximation 6.1 constitutes a popular applied approach, especially considering that maximizing functions is a highly developed enterprise in many branches of applied mathematics.

Example 6.13 *Exponential Sample (Example 6.2 Continued).* Differentiating (6.2) twice with respect to θ gives

$$I_Y(\theta | y) = -n/\theta^2 + 2 \sum y_i/\theta^3.$$

Taking expectations over Y under the exponential specification for y_i gives

$$\begin{aligned} J_Y(\theta) &= -n/\theta^2 + 2E\left(\sum y_i | \theta\right)/\theta^3 \\ &= -n/\theta^2 + 2n\theta/\theta^3 = n/\theta^2. \end{aligned}$$

Substituting the ML estimate $\hat{\theta} = \bar{y}$ for θ gives

$$I_Y(\hat{\theta} | y) = J_Y(\hat{\theta}) = n/\bar{y}^2,$$

whence the large sample variance of $\theta - \hat{\theta}$ is estimated as \bar{y}^2/n .

Example 6.14 *Univariate Normal Sample (Example 6.1 Continued).* For the univariate normal model of Example 6.1, the asymptotic approximation (6.18) is more appropriately applied to $(\mu, \log \sigma^2)$ than to (μ, σ^2) . Differentiating (6.1) twice with respect to μ and $\log \sigma^2$ and substituting ML estimates of the parameters yields

$$I_Y(\hat{\mu}, \log \hat{\sigma}^2 | y) = J_Y(\hat{\mu}, \log \hat{\sigma}^2) = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/2 \end{bmatrix}.$$

Inverting this matrix yields the large sample second moments $\text{Var}(\mu - \hat{\mu}) = \hat{\sigma}^2/n$, $\text{Cov}(\mu - \hat{\mu}, \log \sigma^2 - \log \hat{\sigma}^2) = 0$, and $\text{Var}(\log \sigma^2 - \log \hat{\sigma}^2) = 2/n$, where from Example 6.7, $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = (n-1)s^2/n$.

It is common to summarize evidence about the likely values of a multicomponent θ by “significance levels” rather than by ellipsoids such as (6.19), particularly when the number of components in θ , d , is greater than two. Specifically, for a null value θ_0 of θ , the distance from $\hat{\theta}$ to θ_0 can be calculated as the Wald statistic

$$W(\theta_0, \hat{\theta}) = (\theta_0 - \hat{\theta})^T C^{-1}(\theta_0 - \hat{\theta}),$$

which is the left side of (6.23) evaluated at $\theta = \theta_0$. The associated percentile of the chi-squared distribution on d degrees of freedom is the significance level or p -value of the null value θ_0 :

$$p_C = \Pr\{\chi_d^2 > W(\theta_0, \hat{\theta}) | \theta = \theta_0\},$$

which under Approximation 6.1 does not depend on θ_0 . From the frequentist perspective, the significance level provides the probability that in repeated sampling with $\theta = \theta_0$, the ML estimate will be at least as far from θ_0 as the observed

ML estimate $\hat{\theta}$. A size α (two-sided) test of the null hypothesis $H_0: \theta = \theta_0$ is obtained by rejecting H_0 when the p -value p_C is smaller than α ; common values of α are 0.1, 0.05, and 0.01.

From the Bayesian perspective, p_C gives the large sample posterior probability of the set of θ values with lower posterior density than $\theta_0: \Pr[\theta \in \{\theta | f_Y(\theta | y) < f_Y(\theta_0 | y)\} | y]$; see Box and Tiao (1973) and Rubin (2004, section 2.10) for discussion and examples.

Under Approximation 6.1, an asymptotically equivalent procedure for calculating significance levels is to use the likelihood ratio (LR) statistic to measure the distance between $\hat{\theta}$ and θ_0 ; this yields

$$p_L = \Pr \{ \chi_d^2 > \text{LR}(\theta_0, \hat{\theta}) \},$$

where

$$\text{LR}(\theta_0, \hat{\theta}) = 2 \ln[L_Y(\hat{\theta} | y)/L_Y(\theta_0 | y)] = 2[\ell_Y(\hat{\theta} | y) - \ell_Y(\theta_0 | y)].$$

More generally, suppose $\theta = (\theta_{(1)}, \theta_{(2)})$, and we are interested in evaluating the propriety of a null value of $\theta_{(1)}$, $\theta_{(1)0}$, where the number of components in $\theta_{(1)}$ is q . This situation commonly arises when comparing the fit of two models A and B, which are termed nested because the parameter space for model B is obtained from that for model A by setting $\theta_{(1)}$ to zero. Two asymptotically equivalent approaches derive significance levels corresponding to p_C and p_L as follows:

$$p_C(\theta_{(1)0}) = \Pr \left\{ \chi_q^2 > (\theta_{(1)0} - \hat{\theta}_{(1)})^\top C_{(1)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)}) \right\},$$

where $C_{(1)}$ is the variance–covariance matrix of $\theta_{(1)}$ as in (6.24), and

$$p_L(\theta_{(1)0}) = \Pr \left\{ \chi_q^2 > \text{LR}(\hat{\theta}, \tilde{\theta}) \right\},$$

where

$$\text{LR}(\hat{\theta}, \tilde{\theta}) = 2[\ell_Y(\hat{\theta} | y) - \ell_Y(\tilde{\theta} | y)],$$

and $\tilde{\theta}$ is the value of θ that maximizes $\ell_Y(\theta | y)$ subject to the constraint that $\theta_{(1)} = \theta_{(1)0}$. Level α hypothesis tests reject $H_0: \theta_{(1)} = \theta_{(1)0}$, if the p -value for $\theta_{(1)0}$ is smaller than α .

Example 6.15 Univariate Normal Sample (Example 6.1 Continued). Suppose $\theta = (\mu, \sigma^2)$, $\theta_{(1)} = \mu$, and $\theta_{(2)} = \sigma^2$. To test $H_0: \mu = \mu_0$, the likelihood ratio test statistic is

$$\begin{aligned} \text{LR} &= 2 \left(-(n/2) \ln \{(n-1)s^2/n\} - n/2 + (n/2) \ln s_0^2 + n/2 \right) \\ &= n \ln \left(ns_0^2 / ((n-1)s^2) \right), \end{aligned}$$

where $s_0^2 = n^{-1} \sum_{i=1}^n (y_i - \mu_0)^2 = (n-1)s^2/n + (\bar{y} - \mu_0)^2$. Hence, $\text{LR} = n \ln(1 + t^2/n)$, where $t^2 = n^2(\bar{y} - \mu_0)^2 / \{(n-1)s^2\}$ is, from Example 6.14, the test statistic for H_0 based on the asymptotic variance of $(\mu - \mu_0)$. Asymptotically, $\text{LR} = t^2$ and is chi-squared distributed with $q = 1$ degrees of freedom under H_0 . An exact test is obtained in this case by comparing the statistic t^2 directly with an F distribution on 1 and $n - 1$ degrees of freedom. Such exact small sample tests are rarely available when we apply the likelihood ratio method to data sets with missing values.

Despite the popularity of the approximations discussed here based on asymptotic normality, they can be quite poor in many models with large numbers of parameters or with latent structures. See, for example Rubin and Thayer (1983) and Frumento et al. (2016), which compare asymptotically equivalent approximations and show that the answers can be dramatically different, which can only happen if asymptotic normality fails. Careful Bayesian modeling can yield better results in these settings because the Bayesian approach does not rely on such asymptotic approximations.

6.1.4 Bayes Inference Based on the Full Posterior Distribution

The theory of the previous section assumes large samples, which can yield unsatisfactory inferences when the sample size is too small. One approach to this limitation is to adopt a Bayesian perspective and base inferences on the posterior distribution for a particular choice of prior distribution. To measure uncertainty about the point estimate, the posterior standard deviation is analogous to the frequentist sampling standard deviation of the estimate, and posterior probability intervals (such as the 2.5th to 97.5th percentile of the posterior distribution or the 95% probability interval containing the highest values of the posterior density) are analogous to the frequentist confidence interval. Similarly, the probability associated with the set of values of θ with lower posterior probability than a null value θ_0 replaces the frequentist p -value for testing a null hypothesis.

One issue with this approach is that Bayesian inferences with small samples are more sensitive to the choice of prior distribution than inferences with large samples; frequentist statisticians commonly regard this as the Achilles' heel of the Bayesian approach. However, in cases where prior information is available, better inferences, from a frequentist perspective, can result from formally incorporating this information, using the Bayesian machinery. In situations where prior knowledge is more limited, or where "objective" inferences are sought that are not strongly influenced by prior information, the Bayesian approach with dispersed priors often yields inferences that are better from a frequentist perspective than the large-sample approximations of Section 6.1.3.

In particular, the next two examples show that in problems involving the normal distribution, the Bayesian approach with “noninformative” priors can recover the degrees of freedom corrections and Student t reference distribution arising in small-sample frequentist inference.

For more complex problems, including those involving missing data, the Bayesian approach is attractive because it provides answers in situations where no exact frequentist solutions are available. The Bayesian approach is prescriptive, and no matter how complex the problem, any Bayesian answer, once derived, can be evaluated from a frequentist, repeated sampling, perspective. In particular, the method of multiple imputation discussed in Chapter 10 is based on Bayesian principles, and in general, multiple imputation under a realistic model has excellent frequentist properties.

Example 6.16 *Bayes Inference for a Univariate Normal Sample with Conjugate Prior (Example 6.1 Continued).* For a univariate normal sample with $\theta = (\mu, \sigma^2)$, suppose the following conjugate prior distribution is chosen:

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu | \sigma^2),$$

where

$$\begin{aligned} \sigma^2 / (v_0 \sigma_0^2) &\sim \text{Inv-}\chi^2(v_0), \\ (\mu | \sigma^2) &\sim N(\mu_0, \sigma^2 / \kappa_0), \end{aligned} \tag{6.25}$$

for known v_0 , σ_0^2 , μ , and κ_0 . Here, $\text{Inv-}\chi^2(v_0)$ denotes the inverse of a chi-squared distribution with v_0 degrees of freedom (see Gelman et al. 2013, Appendix A). The prior distribution for μ given σ^2 is normal. The joint prior density for (μ, σ^2) thus has the form:

$$p(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(v_0/2+1)} \exp\left(-\frac{1}{2\sigma^2} [v_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2]\right).$$

It can be shown (Gelman et al. 2013, Section 3.3) that the posterior distribution of θ from data $y = (y_1, \dots, y_n)$ is then

$$p(\mu, \sigma^2 | y) = p(\sigma^2 | y)p(\mu | \sigma^2, y),$$

where the posterior distribution for σ^2 has the scaled inverse chi-squared distribution

$$\begin{aligned} \left[\left(\sigma^2 / v_n \sigma_n^2 \right) | y \right] &\sim \text{Inv-}\chi^2(v_n), \\ v_n &= v_0 + n, \\ v_n \sigma_n^2 &= v_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)^2, \end{aligned} \tag{6.26}$$

and the posterior distribution for μ given σ^2 has the normal distribution

$$\begin{aligned} (\mu \mid \sigma^2, y) &\sim N(\mu_n, \sigma^2/\kappa_n), \\ \kappa_n &= \kappa_0 + n, \\ \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}. \end{aligned} \tag{6.27}$$

Integrating this conditional posterior distribution over the posterior distribution of σ^2 yields the marginal posterior distribution of μ :

$$(\mu \mid y) \sim t\left(\mu_n, \sigma_n^2/\kappa_n, v_n\right), \tag{6.28}$$

the Student's t distribution with center μ_n , squared scale σ_n^2/κ_n , and degrees of freedom v_n .

In the absence of strong prior information, a conventional choice of prior distribution is the Jeffreys' prior distribution:

$$p(\mu, \sigma^2) \propto 1/\sigma^2, \tag{6.29}$$

which is a degenerate special case of (6.25) with $\kappa_0 = 0$, $v_0 = -1$, and $\sigma_0^2 = 0$. Substituting these values in Eqs. (6.26)–(6.28) yields $\mu_n = \bar{y}$, $\kappa_n = n$, $v_n = n - 1$, $\sigma_n^2 = s^2$ and

$$\sigma^2/((n-1)s^2) \mid \tilde{y} \sim \text{Inv-}\chi^2(n-1), \tag{6.30}$$

$$(\mu \mid \sigma^2, y) \sim N(\bar{y}, \sigma^2/n), \tag{6.31}$$

$$(\mu \mid y) \sim t(\bar{y}, s^2/n, n-1). \tag{6.32}$$

In particular, the $100(1-\alpha)\%$ posterior probability interval for μ given y is $\bar{y} \pm t_{1-\alpha/2}s/\sqrt{n}$, where $t_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the Student t distribution with center 0, scale 1, and degrees of freedom $n-1$. This interval is identical to the standard $100(1-\alpha)\%$ confidence interval for μ . Thus, the Bayesian analysis with prior distribution (6.29) recovers the degrees of freedom correction and the t reference distribution of the classical normal-sample frequentist analysis.

Example 6.17 Bayes' Inference for Unweighted and Weighted Multiple Linear Regression (Example 6.10 Continued). The Jeffreys' prior distribution for the normal linear regression model of Example 6.10 is

$$p(\beta_0, \beta_1, \dots, \beta_p, \sigma^2) \propto 1/\sigma^2, \tag{6.33}$$

which places a uniform prior on the location parameters and $\ln \sigma$. The corresponding posterior distribution has the form

$$\sigma^2/((n-p-1)s^2) \mid y \sim \text{Inv-}\chi^2(n-p-1), \tag{6.34}$$

$$(\beta | \sigma^2, y) \sim N_{p+1}(\hat{\beta}, (X^T X)^{-1} \sigma^2), \quad (6.35)$$

$$(\beta | y) \sim t_{p+1}(\hat{\beta}, (X^T X)^{-1} s^2, n - p - 1), \quad (6.36)$$

the multivariate t distribution with center $\hat{\beta}$, scale $(X^T X)^{-1} s^2$, and degrees of freedom $n - p - 1$. Here, $s^2 = \hat{\sigma}^2 / (n - p - 1)$, the residual mean square adjusted for degrees of freedom. Equation (6.36) gives 100(1 - α)% posterior probability intervals for the regression coefficients that match the confidence intervals from normal linear regression theory.

The extension to Bayes' inference for weighted multiple linear regression, where the residual variance of y_i is σ^2/w_i for known w_i , is straightforward; the only change is that $(y_i - \beta x_i)/\sqrt{w_i} \sim_{\text{ind}} N(0, \sigma^2)$. Thus in (6.34)–(6.36), the weighted estimates (6.7) and (6.8) replace the unweighted estimates (6.5) and (6.6), and $X^T W X$ replaces $X^T X$, where W is the diagonal matrix with entries (w_1, \dots, w_n) .

The following two examples are important in practice.

Example 6.18 *Bayes' Inference for a Multinomial Sample (Example 6.3 Continued).* Suppose the data y is a vector of counts (n_1, \dots, n_C) with a multinomial distribution as in Example 6.3. The conjugate prior distribution is the Dirichlet distribution, a multivariate generalization of the beta distribution:

$$p(\pi_1, \dots, \pi_C) \propto \prod_{c=1}^C \pi_c^{\alpha_c - 1}, \quad \pi_c > 0, \sum_{c=1}^C \pi_c = 1. \quad (6.37)$$

Combining this prior distribution with the likelihood (6.3) yields the posterior distribution as Dirichlet with parameters $\{\tilde{n}_c + \alpha_c\}$:

$$p(\pi_1, \dots, \pi_C | y) \propto \prod_{c=1}^C \pi_c^{n_c + \alpha_c - 1}, \quad \pi_c > 0, \sum_{c=1}^C \pi_c = 1. \quad (6.38)$$

By properties of the Dirichlet distribution (Gelman et al. 2013, Appendix A), the posterior mean of π_c is $(n_c + \alpha_c)/(n + \alpha_+)$, where $n = \sum_{c=1}^C n_c$, which equals the ML estimate when $\alpha_c = 0$ for all c . Alternative relatively diffused prior distributions are given by $\alpha_c = 1$ for all c , yielding the uniform distribution, or $\alpha_c = 0.5$ for all c , which yields the Jeffreys' prior distribution for this problem.

Example 6.19 *Bayes Inference for a Multivariate Normal Sample (Example 6.4 Continued).* The noninformative Jeffreys' prior distribution for a multivariate normal sample with $\theta = (\mu, \Sigma)$ is

$$p(\mu, \Sigma) \propto |\Sigma|^{-(K+1)/2}, \quad (6.39)$$

which reduces to (6.29) when $K = 1$. The corresponding posterior distribution can be written as follows:

$$\begin{aligned} (\Sigma/(n-1) | y) &\sim \text{Inv-Wishart}_{n-1}(S^{-1}), \\ (\mu | \Sigma, y) &\sim N_K(\bar{y}, \Sigma/n), \end{aligned} \quad (6.40)$$

where $\text{Inv-Wishart}_{n-1}(S^{-1})$ denotes the inverse Wishart distribution with $n - 1$ degrees of freedom and scale matrix S (see Gelman et al. 2013, Appendix A). Equation (6.40) implies that the marginal posterior distribution of μ is multivariate t with center \bar{y} , scale matrix S/n , and degrees of freedom $n - 1$.

6.1.5 Simulating Posterior Distributions

The application of Bayes' methods to more complex problems used to be constrained by the numerical difficulties involved when computing the posterior distribution of parameters, particularly when θ was high dimensional. For example if $\theta = (\theta_1, \theta_2)$, then the posterior distribution of θ_1 is

$$p(\theta_1 | y) = \int p(\theta) L_Y(\theta | y) d\theta_2 / \int p(\theta) L_Y(\theta | y) d\theta,$$

which involves high-dimensional integration when θ_2 has many components. These problems have been reduced by simulation methods that take draws from the posterior distribution of θ , rather than trying to compute the distribution analytically. These draws can be used to estimate characteristics of the posterior distribution of interest. For example, the mean and variance of the posterior distribution of scalar θ_1 can be estimated as the sample mean, and variance of D draws $(\theta_1^{(d)}, d = 1, \dots, D)$. If the posterior distribution is far from normal, 95% central probability intervals for θ_1 can be estimated as the 2.5th to 97.5th percentiles of the empirical distribution of the draws $\{\theta_1^{(d)}\}$.

The transformation Property 6.1 for ML estimates has a direct analog for draws from the posterior distribution. We denote the analog as Property 6.1B, where B stands for "Bayes."

Property 6.1B Let $g(\theta)$ be a function of the parameter θ , and let $\theta^{(d)}$ be the d th draw from the posterior distribution of θ , $d = 1, \dots, D$. Then, $g(\theta^{(d)})$ is a draw from the posterior distribution of $g(\theta)$.

This property is useful in applications of Bayes simulation methods to incomplete-data problems, as discussed in Section 7.3 and Chapter 10.

Example 6.20 *Bayes' Inference for Multiple Linear Regression (Example 6.17 Continued).* Draws $\{(\beta^{(d)}, \sigma^{(d)}), d = 1, \dots, D\}$ from the posterior

distribution of (β, σ) for the normal regression model data of Example 6.17 with prior distribution (6.33) are readily obtained from (6.34) and (6.35) as follows:

1. For $d = 1, \dots, D$, draw χ_{n-p-1}^2 from a chi-squared distribution with $n - p - 1$ degrees of freedom, and set

$$\sigma^{(d)2} = (n - p - 1)s^2 / \chi_{n-p-1}^2. \quad (6.41)$$

2. Draw $p + 1$ standard normal deviates, $z = (z_0, z_1, \dots, z_p)^T$, $z_i \sim N(0, 1)$, $i = 0, 1, \dots, p$, and set

$$\beta^{(d)} = \hat{\beta} + A^T z \sigma^{(d)}, \quad (6.42)$$

where A is an upper triangular $(p \times p)$ Cholesky factor of $(X^T X)^{-1}$, such that $A^T A = (X^T X)^{-1}$. Draws are not needed for inference about the regression coefficients themselves, which have a t distribution. However, they are often needed to simulate the posterior distribution of nonlinear functions of the parameters, using Property 6.1B. For example, a draw from the posterior distribution of $\lambda = \beta_1/\beta_2$ is simply $\lambda^{(d)} = \beta_1^{(d)}/\beta_2^{(d)}$. The draws (6.41) and (6.42) play an important role for simulating posterior distributions for normal missing-data problems, discussed later.

Example 6.21 *Bayes Inference for a Multinomial Sample (Example 6.18 Continued).* Draws $\{\pi_c^{(d)}\}$ from the Dirichlet posterior distribution (6.38) of $\{\pi_c\}$ under the multinomial model of Example 6.18 can be obtained by generating independent chi-squared deviates $\{\chi_{2(n_c + \alpha_c)}^2\}$ for $c = 1, \dots, C$, and setting

$$\pi_c^{(d)} = \chi_{2(n_c + \alpha_c)}^2 / \sum_{j=1}^C \chi_{2(n_j + \alpha_j)}^2. \quad (6.43)$$

Often chi-squared random variables, and the associated t distribution, are defined with integer degrees of freedom. This restriction is unnecessary because a more general form of (6.43) replaces the $\chi_{2(n_c + \alpha_c)}^2$ random variables by a standard (scale parameter = 1) gamma distribution with parameters $n_c + \alpha_c$, with density proportional to $f(x | \alpha_c, n_c) = x^{n_c + \alpha_c - 1} \exp(-x)$. See, for example, Gelman et al. (2013, Appendix A). For the special case of $C = 2$, the multinomial sample is a binomial sample, and the Dirichlet prior and posterior distributions become beta prior and posterior distributions.

Example 6.22 *Bayes' Inference for a Multivariate Normal Sample (Example 6.19 Continued).* The posterior distribution of $\theta = (\mu, \Sigma)$ for a multivariate normal sample with prior distribution (6.39) is given by (6.40). A draw from this distribution is obtained by first drawing $C^{(d)}$ from $\text{Inv-Wishart}_{n-1}(S^{-1})$, setting

$\Sigma^{(d)} = (n - 1)C^{(d)}$, and then drawing $\mu^{(d)} = \bar{y} + A^{(d)\top}z$, where $z = (z_1, \dots, z_K)^\top$ is a vector of independent $N(0, 1)$ draws, and $A^{(d)}$ is an upper triangular Cholesky factor such that $A^{(d)\top}A^{(d)} = \Sigma^{(d)}/n$.

The inverse-Wishart draw $C^{(d)}$ can be obtained by forming an upper triangular matrix B with elements

$$b_{jj} \sim \sqrt{\chi^2_{n-j}}, \quad b_{jk} \sim N(0, 1), \quad j < k, \quad (6.44)$$

and drawing

$$C^{(d)} = (B^\top)^{-1}A, \quad (6.45)$$

where A is the Cholesky factor of S^{-1} , that is $A^\top A = S^{-1}$. These results follow from the Bartlett decomposition of the Wishart distribution (e.g., Muirhead 1982).

6.2 Likelihood-Based Inference with Incomplete Data

In one high-level sense, there is no formal difference between ML or Bayes inference for incomplete data and ML or Bayes inference for complete data: The likelihood for the parameters based on the incomplete data is derived, ML estimates are found by solving the likelihood equation, and the posterior distribution is obtained by incorporating a prior distribution and performing the necessary integrations. Asymptotic standard errors obtained from the information matrix are somewhat more questionable with missing data, however, because the observed data do not generally constitute an iid sample, and simple results that imply the large sample normality of the likelihood function do not immediately apply. Other complications arise from dealing with the process that creates missing data. We will be somewhat imprecise in our treatment of these complications, to keep the notation simple. Rubin (1976a) gives a mathematically precise treatment, which also encompasses frequentist approaches that are not based on the likelihood; Mealli and Rubin (2015) provides an updated version for likelihood-based inference.

Let $Y = (y_{ij})$, $i = 1, \dots, n$, $j = 1, \dots, K$ denote the data matrix if there were no missing values, with n units and K variables, with $y_{ij} \in \Omega_{ij}$, its sample space. Let $M = (m_{ij})$ denote the fully observed $(n \times K)$ matrix of binary missingness indicators with $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. When $m_{ij} = 1$, $y_{ij} = *$, indicating that y_{ij} can take any value in Ω_{ij} . We can model the density of the joint distribution of Y and M using the “selection model” factorization (Little and Rubin 2002)

$$p(Y = y, M = m | \theta, \psi) = f_Y(y | \theta)f_{M|Y}(m | y, \psi), \quad (6.46)$$

where θ is the parameter vector governing the data model, and ψ is the parameter vector governing the model for the missingness mechanism. The observed value m of M effects a partition of $y = (y_{(0)}, y_{(1)})$, where $y_{(0)} = [y_{ij} : m_{ij} = 0]$ is the observed part of y and $y_{(1)} = [y_{ij} : m_{ij} = 1]$ is the missing part of y .²

The full likelihood based on the observed values $(y_{(0)}, m)$ and the assumed model (6.46) is defined to be

$$L_{\text{full}}(\theta, \psi | y_{(0)}, m) = \int f_Y(y_{(0)}, y_{(1)} | \theta) f_{M|Y}(m | y_{(0)}, y_{(1)}, \psi) dy_{(1)}, \quad (6.47)$$

considered as a function of the parameters (θ, ϕ) . The likelihood of θ *ignoring the missingness mechanism* is defined to be

$$L_{\text{ign}}(\theta | y_{(0)}) = \int f_Y(y_{(0)}, y_{(1)} | \theta) dy_{(1)}, \quad (6.48)$$

which does not involve the model for M . The term “ignorable likelihood” is sometimes used for Eq. (6.48), hence the notation L_{ign} .

As discussed in detail in Chapter 15, modeling the joint distribution of M and Y is often challenging. Many approaches to missing data do not model M , and (implicitly or explicitly) base inference about θ on the ignorable likelihood, Eq. (6.48). It is thus important to consider under what conditions inferences about θ can be based on this simpler likelihood. The following definition addresses this question in general terms:

Definition 6.4 Let \tilde{m} and $\tilde{y}_{(0)}$ denote particular realizations of $(m, y_{(0)})$. The missingness mechanism is called *ignorable* if inferences for θ based on the ignorable likelihood equation (6.48) evaluated at $m = \tilde{m}, y_{(0)} = \tilde{y}_{(0)}$ are the same as inferences for θ based on the full likelihood equation (6.47) evaluated at $m = \tilde{m}, y_{(0)} = \tilde{y}_{(0)}$.

The conditions for ignoring the missingness mechanism depend on whether the inferences are direct likelihood, Bayes or frequentist. We consider each of these forms of inference in turn.

Direct likelihood inference, as discussed in Section 6.1.2, refers to inference based solely on likelihood ratios for pairs of values of the parameters, with the data fixed at their observed values. The missingness mechanism can be ignored for direct likelihood if the likelihood ratio based on the ignorable likelihood is the same as the ratio based on the full model. The following definition states this more precisely:

Definition 6.4A For the model defined by Eq. (6.46), the missingness mechanism is called ignorable for direct likelihood inference at $(\tilde{m}, \tilde{y}_{(0)})$ if the

likelihood ratio for two values θ and θ^* is the same whether based on the full likelihood or the ignorable likelihood. That is

$$\frac{L_{\text{full}}(\theta, \psi | \tilde{y}_{(0)}, \tilde{m})}{L_{\text{full}}(\theta^*, \psi | \tilde{y}_{(0)}, \tilde{m})} = \frac{L_{\text{ign}}(\theta | \tilde{y}_{(0)})}{L_{\text{ign}}(\theta^* | \tilde{y}_{(0)})} \quad \text{for all } \theta, \theta^*, \psi. \quad (6.49)$$

Theorem 6.1A *The missingness mechanism is ignorable for direct likelihood inference $(\tilde{m}, \tilde{y}_{(0)})$ if the following two conditions hold:*

- (a) *Parameter distinctness: The parameters θ and ψ are distinct, in the sense that the joint parameter space (θ, ψ) , say $\Omega_{\theta, \psi}$, is the product of the parameter space Ω_θ of θ and the parameter space Ω_ψ of ψ , that is $\Omega_{\theta, \psi} = \Omega_\theta \times \Omega_\psi$.*
- (b) *Factorization of the full likelihood: The full likelihood, Eq. (6.47), with $(y_0, m) = (\tilde{y}_0, \tilde{m})$ factors as*

$$L_{\text{full}}(\theta, \psi | \tilde{y}_{(0)}, \tilde{m}) = L_{\text{ign}}(\theta | \tilde{y}_{(0)}) \times L_{\text{rest}}(\psi | \tilde{y}_{(0)}, \tilde{m}) \quad \text{for all } \theta, \psi \in \Omega_{\theta, \psi}. \quad (6.50)$$

Theorem 6.1A follows directly after substituting Eq. (6.50) into the left side of Eq. (6.49). The distinctness condition ensures that each value of $\psi \in \Omega_\psi$ is compatible with different values of $\theta \in \Omega_\theta$, such as θ and θ^* in Eq. (6.49).

A sufficient condition for the likelihood to factor as in Eq. (6.50) is that the missing data are missing at random (MAR) at $(\tilde{m}, \tilde{y}_{(0)})$, which is defined at an intuitive level in Section 1.3. Formally, the missing data are MAR at $(\tilde{m}, \tilde{y}_{(0)})$ when the distribution function of M evaluated at the observed values $(\tilde{y}_{(0)}, \tilde{m})$ of $(y_{(0)}, m)$ does not depend on the missing values $y_{(1)}$, that is

$$f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, y_{(1)}, \psi) = f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, y_{(1)}^*, \psi) \quad \text{for all } y_{(1)}, y_{(1)}^*, \psi. \quad (6.51)$$

Under Eq. (6.51),

$$\begin{aligned} f(\tilde{y}_{(0)}, \tilde{m} | \theta, \psi) &= f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, \psi) \times \int f_Y(\tilde{y}_{(0)}, y_{(1)} | \theta) dy_{(1)} \\ &= f_{M|Y}(\tilde{m} | \tilde{y}_{(0)}, \psi) \times f_Y(\tilde{y}_{(0)} | \theta), \end{aligned}$$

yielding the factored likelihood (6.50). Thus, we have the following corollary:

Corollary 6.1A *If the missing data are MAR at $(\tilde{m}, \tilde{y}_{(0)})$, and θ and ψ are distinct, the missingness mechanism is ignorable for likelihood inference.*

For Bayesian inference under the full model Eq. (6.46) for Y and M , the full likelihood (6.49) is combined with a prior distribution $p(\theta, \psi)$ for θ and ψ

$$p(\theta, \psi | \tilde{y}_{(0)}, \tilde{m}) \propto p(\theta, \psi) \times L_{\text{full}}(\theta, \psi | \tilde{y}_{(0)}, \tilde{m}). \quad (6.52)$$

Bayesian inference ignoring the missingness mechanism combines the ignorable likelihood (6.48) with a prior distribution for θ alone, that is

$$p(\theta | \tilde{y}_{(0)}) \propto p(\theta) \times L_{\text{ign}}(\theta | \tilde{y}_{(0)}). \quad (6.53)$$

We now provide parallel expressions to Definition 6.4A, Theorem 6.1A, and Corollary 6.1A for Bayesian inference:

Definition 6.4B The missingness mechanism is called ignorable for Bayesian inference at $(\tilde{m}, \tilde{y}_{(0)})$ if the posterior distribution for θ based on Eq. (6.53) is the same as the posterior distribution for θ based on Eq. (6.52).

Theorem 6.1B *The posterior distribution for θ based on Eq. (6.52) is the same as the posterior distribution for θ based on Eq. (6.53) if: (a) the parameters θ and ψ are a priori independent, that is, the prior distribution has the form*

$$p(\theta, \psi) = p(\theta)p(\psi); \quad (6.54)$$

and (b) the full likelihood evaluated at \tilde{m} and \tilde{y}_0 factors as in Eq. (6.50).

Theorem 6.1B follows because under conditions (a) and (b),

$$p(\theta, \psi | \tilde{y}_{(0)}, \tilde{m}) \propto \{p(\theta)L_{\text{ign}}(\theta | \tilde{y}_{(0)})\} \times \{p(\psi)L_{\text{rest}}(\psi | \tilde{y}_{(0)}, \tilde{m})\}$$

so the posterior distribution of θ based on Eq. (6.52) reduces to the posterior distribution given by Eq. (6.53). As for direct likelihood inference, MAR at $(\tilde{m}, \tilde{y}_{(0)})$ is sufficient for condition (b), that the likelihood factors as in Eq. (6.50). Thus,

Corollary 6.1B *If the missing data are MAR at $(\tilde{m}, \tilde{y}_{(0)})$, and the parameters θ and ψ are a priori independent, then the missingness mechanism is ignorable for Bayesian inference.*

Note that the *a priori* independence condition, Eq. (6.54), is more stringent than the distinctness condition for direct likelihood inference because parameters with distinct parameter spaces might have dependent prior distributions.³

In order to ignore the missingness mechanism for asymptotic frequentist likelihood inference, the factorization Eq. (6.50) in general needs to be valid for values of the observed data in repeated sampling. That is, we require that

$$\begin{aligned} L_{\text{full}}(\theta, \psi | y_{(0)}, m) &= L_{\text{ign}}(\theta | y_{(0)}) \times L_{\text{rest}}(\psi | y_{(0)}, m) \\ \text{for all } y_{(0)}, m \text{ and } \theta, \psi \in \Omega_{\theta, \psi}. \end{aligned} \quad (6.55)$$

For this form of inference, a sufficient condition for ignoring the missingness mechanism is a parameter distinctness, as in Definition 6.4, and that the missing data are missing always at random (MAAR), that is

$$f_{M|Y}(m \mid y_{(0)}, y_{(1)}, \psi) = f_{M|Y}(m \mid y_{(0)}, y_{(1)}^*, \psi) \quad \text{for all } m, y_{(0)}, y_{(1)}, y_{(1)}^*, \psi. \quad (6.56)$$

For ease of presentation, in what follows, we discuss conditions for ignoring the missingness mechanism for direct likelihood and Bayes inference. However, we note that parallel conditions can be developed for frequentist likelihood inference, by requiring that the relevant equations apply for values of $(m, y_{(0)})$, other than those observed that could arise in repeated sampling.

Example 6.23 Incomplete Exponential Sample. Suppose we have an incomplete univariate exponential sample with $y_{(0)} = (y_1, \dots, y_r)^T$ observed and $y_{(1)} = (y_{r+1}, \dots, y_n)^T$ missing. Thus, $m = (m_1, \dots, m_n)^T$, where $m_i = 0$, $i = 1, \dots, r$ and $m_i = 1$, $i = r+1, \dots, n$. As in Example 6.2,

$$f_Y(y \mid \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^n \frac{y_i}{\theta}\right).$$

The likelihood ignoring the missingness mechanism is

$$L_{\text{ign}}(\theta \mid y_{(0)}) = \theta^{-r} \exp\left(-\sum_{i=1}^r \frac{y_i}{\theta}\right). \quad (6.57)$$

Suppose that each unit is observed with probability ψ that does not depend on Y , so that Eq. (6.51) holds. Then,

$$f_{M|Y}(m \mid y, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r},$$

and

$$f(y_{(0)}, m \mid \theta, \psi) = \frac{n!}{r!(n-r)!} \psi^r (1-\psi)^{n-r} \theta^{-r} \exp\left(-\sum_{i=1}^r \frac{y_i}{\theta}\right).$$

Because the missing data are MAR, if ψ and θ are distinct, then likelihood-based inferences about θ can be based on the ignorable likelihood, Eq. (6.57). In particular, the ML estimate of θ is simply $\sum_{i=1}^r y_i/r$, the mean of the observed values of Y .

Suppose, instead, that the incomplete data are created by censoring at some known censoring point c , so that only values less than c are observed. Then,

$$f_{M|Y}(m \mid y, \psi) = \prod_{i=1}^n f(m_i \mid y_i, \psi),$$

where

$$f(m_i | y_i, \psi) = \begin{cases} 1, & \text{if } m_i = 1 \text{ and } y_i \geq c, \text{ or } m_i = 0 \text{ and } y_i < c, \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} L_{\text{full}}(\theta | y_{(0)}, m) &= \prod_{i=1}^r f_Y(y_i | \theta) \Pr(y_i < c | y_i, \theta) \times \prod_{i=r+1}^n \Pr(y_i \geq c | \theta) \\ &= \theta^{-r} \exp\left(-\sum_1^r \frac{y_i}{\theta}\right) \times \exp\left(-\frac{(n-r)c}{\theta}\right), \end{aligned} \quad (6.58)$$

because $\Pr(y_i < c | y_i, \theta) = 1$ for respondents and $\Pr(y_i \geq c | \theta) = \exp(-c/\theta)$ for nonrespondents, using the properties of the exponential distribution. In this case, the missingness mechanism is not ignorable for likelihood inference, and the correct likelihood equation (6.58) differs from ignorable likelihood equation (6.57). Maximizing Eq. (6.58) with respect to θ gives the ML estimate $\hat{\theta} = (\sum_{i=1}^r y_i + (n-r)c)/r$, which can be compared with the (incorrect) ignorable ML estimate $\sum_{i=1}^r y_i/r$. The inflation of the sample mean in this expression reflects the censoring of the missing values.

Example 6.24 Bivariate Normal Sample with One Variable Subject to Missingness. Consider a bivariate normal sample, as in Example 6.9, but with the values y_{i2} of the second variable missing for $i = (r+1), \dots, n$. We thus have a monotone pattern with two variables. The loglikelihood ignoring the missingness mechanism is

$$\begin{aligned} \ell_{\text{ign}}(\mu, \Sigma | y_{(0)}) &= \log(L_{\text{ign}}(\mu, \Sigma | y_{(0)})) = -\frac{1}{2}r \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^r (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)^T \\ &\quad - \frac{1}{2}(n-r) \ln \sigma_{11} - \frac{1}{2} \sum_{i=r+1}^n \frac{(y_{i1} - \mu_1)^2}{\sigma_{11}}. \end{aligned} \quad (6.59)$$

This loglikelihood is appropriate for inferences provided the conditional distribution of M (and in particular, the probability that y_{i2} is missing) does not depend on the values of y_{i2} (although it may depend on the values of y_{i1}), and $\theta = (\mu, \Sigma)$ is distinct from ψ , the parameters of the missingness mechanism. Under these conditions, ML estimates of μ and Σ can be found by maximizing (6.59). For Bayes inference, if these conditions hold and the prior distribution for (μ, Σ, ψ) has the form $p(\mu, \Sigma)p(\psi)$, then the joint posterior distribution of μ and Σ is proportional to the product of $p(\mu, \Sigma)$ and $L_{\text{ign}}(\mu, \Sigma | y_{(0)})$ from Eq. (6.59). A simple approach to these analyses based on factoring the likelihood is described in the next chapter.

With “random effects” models, there is a subtle interplay between the assumptions of MAR and distinctness (or *a priori* independence), depending on the definition of the hypothetical complete data. The following example illustrates aspects of this subtlety.

Example 6.25 One-Way ANOVA with Missing Values, When Missingness Depends on the Unobserved Group Means. Consider a one-way random effects ANOVA with I groups, data $X = \{x_{ij} : i = 1, \dots, I; j = 1, \dots, n_i\}$, and model

$$(x_{ij} | \mu_i, \theta) \sim_{\text{ind}} N(\mu_i, \sigma^2), \quad (6.60)$$

$$(\mu_i | \theta) \sim_{\text{ind}} N(\mu, \tau^2), \quad (6.61)$$

where $\theta = (\mu, \sigma^2, \tau^2)$ are fixed unknown parameters; that is the unobserved mean for group i , μ_i , is assumed to be sampled from the normal distribution (6.61), for $i = 1, \dots, I$. Suppose that some data values x_{ij} are missing, and let $x_{(0)}$ and $x_{(1)}$ denote the observed and missing values of X , respectively. Suppose the missingness mechanism depends on the unobserved random variables $\{\mu_i\}$:

$$\Pr(m_{ij} = 1 | x, \mu_i, \psi) \equiv \pi(\mu_i, \psi) = \exp(\psi_0 + \psi_1 \mu_i) / (1 + \exp(\psi_0 + \psi_1 \mu_i)). \quad (6.62)$$

For likelihood inference, unknowns without a distribution are parameters, and unknowns with a distribution are missing data; see Section 6.3. Thus, the complete data must be defined to include the unobserved group means μ_i , that is

$$y = (x, \{\mu_i\}), y_{(0)} = x_{(0)}, \text{ and } y_{(1)} = (x_{(1)}, \{\mu_i\}),$$

and the missingness mechanism (6.62) is nonignorable because the missing data are not MAR. Likelihood inference for the model defined by Eqs. (6.60)–(6.62) must be based on the full likelihood, Eq. (6.47). Suppose that r_i values of X are observed in group i , $r = \sum_{i=1}^I r_i$, and let $\bar{x}_{(0)i}$ denote the mean of these observed values and $s_{(0)i}^2$ their sample variance with denominator r_i . The full likelihood is then

$$L_{\text{full}}(\theta, \psi | y_{(0)}, m) = \sigma^{-r} \tau^{-I} \prod_{i=1}^I \int \pi(\mu_i, \psi)^{r_i} (1 - \pi(\mu_i, \psi))^{n_i - r_i} \exp \left(-r_i \left(s_{(0)i}^2 + (\bar{x}_{(0)i} - \mu_i)^2 \right) / (2\sigma^2) - (\mu_i - \mu)^2 / (2\tau^2) \right) d\mu_i.$$

A common alternative model to Eqs. (6.60) and (6.61) is the fixed-effects ANOVA model

$$x_{ij} | \mu_i, \sigma^2 \sim_{\text{ind}} N(\mu_i, \sigma^2), \quad (6.63)$$

where all components of $\theta^* = (\{\mu_i\}, \sigma^2)$ are regarded as the fixed unknown parameters. The data then do not include $\{\mu_i\}$ because they are fixed parameters, and the complete data are defined as

$$y = (y_{(0)}, y_{(1)}), \quad \text{where } y_{(0)} = x_{(0)} \text{ and } y_{(1)} = x_{(1)}.$$

The full likelihood with missingness mechanism (6.62) is then

$$\begin{aligned} L_{\text{full}}(\theta^*, \psi | y_{(0)}, m) \\ = \sigma^{-r} \prod_{i=1}^I \exp(-r_i(\bar{x}_{(0)i} - \mu_i)^2 / (2\sigma^2)) \pi(\mu_i, \psi)^{r_i} (1 - \pi(\mu_i, \psi))^{n_i - r_i}, \end{aligned} \quad (6.64)$$

and the likelihood ignoring the missingness mechanism is then

$$L_{\text{ign}}(\theta^* | y_{(0)}) = \sigma^{-r} \prod_{i=1}^I \exp(-r_i(\bar{x}_{(0)i} - \mu_i)^2 / (2\sigma^2)). \quad (6.65)$$

The missingness mechanism (6.62) is now MAR because (unlike in the random effects model) it does not depend on missing data. However, the distinctness condition in Definition 6.4 is violated because the models (6.62) and (6.63) both involve the parameters $\{\mu_i\}$. Hence, the missingness mechanism is nonignorable for likelihood inferences about θ^* . Likelihood or Bayes inference ignoring the missingness mechanism, based on Eq. (6.65), is strictly speaking incorrect because a relevant part of the full likelihood equation (6.64) is being ignored. From the frequentist perspective, estimators based on Eq. (6.65) can be asymptotically unbiased despite the violation of the distinctness condition, but they are not generally efficient.

The sufficient conditions for ignoring the mechanism for likelihood inference discussed previously are for the full parameter vector θ . Little et al. (2016a) propose the following definitions of partially MAR as well as ignorability for direct likelihood inferences for a subvector θ_1 of the parameters θ in a model:

Definition 6.5 Write $\theta = (\theta_1, \theta_2)$, where θ_1 and θ_2 are subvectors of the components of the model for the data X in Eq. (6.46). The data are partially MAR for direct likelihood inference about θ_1 , denoted P - MAR(θ_1), if the likelihood (6.47) can be factored as

$$L_{\text{full}}(\theta_1, \theta_2, \psi | \tilde{y}_{(0)}, \tilde{m}) = L_1(\theta_1 | \tilde{y}_{(0)}) \times L_{\text{rest}}(\theta_2, \psi | \tilde{y}_{(0)}, \tilde{m}) \quad \text{for all } \theta_1, \theta_2, \psi, \quad (6.66)$$

where $L_1(\theta_1 | \tilde{y}_{(0)})$ does not involve the model for the missingness mechanism, and $L_{\text{rest}}(\theta_2, \psi | \tilde{y}_{(0)}, \tilde{m})$ does not involve the parameters θ_1 .

Definition 6.6 The data are ignorable for direct likelihood inference about θ_1 , denoted $\text{IGN}(\theta_1)$, if (i) the missingness mechanism is $\text{P-MAR}(\theta_1)$ and (ii) θ_1 and (θ_2, ϕ) are distinct sets of parameters, in the sense given in Theorem 6.1A.

When the data are $\text{IGN}(\theta_1)$, likelihood inference for θ_1 based on $L_1(\theta_1 \mid \tilde{y}_{(0)})$ is the same as inference under the full likelihood (6.66) because

$$\begin{aligned}\frac{L_{\text{full}}(\theta_1^*, \theta_2, \psi \mid \tilde{y}_{(0)}, \tilde{m})}{L_{\text{full}}(\theta_1^{**}, \theta_2, \psi \mid \tilde{y}_{(0)}, \tilde{m})} &= \frac{L_1(\theta_1^* \mid \tilde{y}_{(0)})}{L_1(\theta_1^{**} \mid \tilde{y}_{(0)})} \times \frac{L_{\text{rest}}(\theta_2, \psi \mid \tilde{y}_{(0)}, \tilde{m})}{L_{\text{rest}}(\theta_2, \psi \mid \tilde{y}_{(0)}, \tilde{m})} \\ &= \frac{L_1(\theta_1^* \mid \tilde{y}_{(0)})}{L_1(\theta_1^{**} \mid \tilde{y}_{(0)})} \quad \text{for all } \theta_1^*, \theta_1^{**}, \theta_2, \phi.\end{aligned}$$

Similarly, for Bayesian inference, if θ_1 and (θ_2, ψ) are *a priori* independent, the posterior distribution of θ_1 is proportional to $\pi_1(\theta_1) \times L_1(\theta_1 \mid \tilde{y}_{(0)})$, where $\pi_1(\theta_1)$ is the prior distribution of θ_1 .

If the missingness mechanism is $\text{P-MAR}(\theta_1)$, but θ_1 and (θ_2, ϕ) are not distinct parameters, partial likelihood inference based on $L_1(\theta_1 \mid \tilde{y}_{(0)})$ is valid in certain asymptotic frequentist senses, and although not fully efficient, might still be entertained to avoid the additional assumptions involved in modeling the missingness mechanism. The partial likelihood $L_1(\theta_1 \mid \tilde{y}_{(0)})$ combined with a prior distribution for θ_1 yields a form of “pseudo-Bayesian” inference, which is not fully Bayes, but again avoids a model for the missingness mechanism. This approach to inference has been proposed and discussed in other contexts (for example Sinha and Ibrahim 2003; Ventura et al. 2009; Pauli et al. 2011).

Example 6.26 Regression Where Missingness Depends on the Covariates. Suppose the complete data are a random sample (y_i, z_i) , $i = 1, \dots, n$ of values of variables Y and Z , each of which may be vectors, and interest concerns the parameters θ_1 of the regression of Y on Z . Let $(y_{i(1)}, y_{i(0)})$ and $(z_{i(1)}, z_{i(0)})$ denote, respectively, the observed and missing components of y_i and z_i , $i = 1, \dots, n$, and let $m_i = (m_i^{(Y)}, m_i^{(Z)})$, where $m_i^{(Y)}$ and $m_i^{(Z)}$ are vectors of indicators for whether the components of y_i and z_i are missing, respectively. Suppose that for $i = 1, \dots, r$, z_i is fully observed, and at least one component of y_i is observed. For the remaining units, $i = r + 1, \dots, n$, y_i is entirely missing, and the pattern of missing data for z_i is arbitrary, but with at least one component of z_i missing. We assume (y_i, z_i, m_i) are iid across units, with

$$\begin{aligned}f_{Y,Z,M}(y_i, z_i, m_i \mid \theta_1, \theta_2, \psi) &= f_{Y|Z}(y_i \mid z_i, \theta_1) f_Z(z_i \mid \theta_2) \\ f_{M|Y,Z}(m_i \mid z_i, y_i, \psi),\end{aligned}\tag{6.67}$$

and also assume that the missingness mechanism depends on the covariates Z but not on the outcomes Y , that is

$$\begin{aligned} f_{M|Y,Z}(\tilde{m}_i | \tilde{z}_{i(1)}, z_{i(0)}, y_i, \psi) \\ = f_{M|Z}(\tilde{m}_i | \tilde{z}_{i(1)}, z_{i(0)}, y_i^*, \psi) \quad \text{for all } z_{i(0)}, y_i, y_i^*, i = 1, \dots, n. \end{aligned}$$

This missingness mechanism is MNAR because missingness for $i = r+1, \dots, n$ can depend on missing components of Z . The full likelihood factors as

$$L(\theta_1, \theta_2, \psi) = L_1(\theta_1 | \tilde{Y}_{(1)}, \tilde{Z}_{(1)}) \times L_{\text{rest}}(\theta_2, \psi),$$

where

$$L_1(\theta_1 | \tilde{Y}_{(0)}, \tilde{Z}_{(0)}) = \prod_{i=1}^r \int f_{Y|Z}(\tilde{y}_{i(0)}, y_{i(1)} | \tilde{z}_i, \theta_1) dy_{i(1)},$$

and

$$\begin{aligned} L_{\text{rest}}(\theta_2, \psi) &= \prod_{i=1}^r f_Z(\tilde{z}_i | \theta_2) f_{M|Z}(\tilde{m}_i | \tilde{z}_i, \phi) \times \prod_{i=m+1}^n f_Z(\tilde{z}_{i(1)}, \tilde{z}_{i(0)} | \theta_2) \\ &\quad \int f_{M|Z}(\tilde{m}_i | \tilde{z}_{i(0)}, z_{i(1)}, \psi) dz_{i(1)}. \end{aligned}$$

Hence, the data are P-MAR(θ_1) and IGN(θ_1) if θ_1 and (θ_2, ϕ) are distinct parameters. Thus, valid inferences for θ_1 can be based on the data in the first m cases without modeling the missingness mechanism, though there may be information about (θ_2, ϕ) lost in the discarded data. For the special case where the first m units have no missing values of Y , this is complete-case analysis, which is known to be valid in a frequentist sense, as discussed in Example 3.3.

6.3 A Generally Flawed Alternative to Maximum Likelihood: Maximizing over the Parameters and the Missing Data

6.3.1 The Method

A different approach to handling incomplete data, occasionally encountered in the literature, is to treat the missing data as parameters and to maximize the complete data likelihood over both the missing data and parameters. That is, let

$$L_{\text{misp}}(\theta, y_{(1)} | \tilde{y}_{(0)}) = f_Y(\tilde{y}_{(0)}, y_{(1)} | \theta) \tag{6.68}$$

be regarded as a function of $(\theta, y_{(1)})$ for fixed $\tilde{y}_{(0)}$, and estimate θ by maximizing $L_{\text{misp}}(\theta, y_{(1)} \mid \tilde{y}_{(0)})$ over both θ and $y_{(1)}$. When the missing data are not MAR, or θ is not distinct from ψ , θ would be estimated in this approach by maximizing

$$\begin{aligned} L_{\text{misp}}(\theta, \psi, y_{(1)} \mid \tilde{y}_{(0)}, \tilde{m}) &= L_{\text{full}}(\theta, \psi \mid \tilde{y}_{(0)}, y_{(1)}, \tilde{m}) \\ &= f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) f_{M|Y}(\tilde{m} \mid \tilde{y}_{(0)}, y_{(1)}, \psi) \end{aligned} \quad (6.69)$$

over $(\theta, \psi, y_{(1)})$. Although this approach can be useful in particular problems, such as described in Chapter 2, it is not a generally valid approach to the analysis of incomplete data. The optimal frequentist properties of ML estimation require that the number of parameters increases at a slow enough rate relative to the sample size. This condition implies that maximizing L_{misp} over $(\theta, \psi, y_{(1)})$ is only generally valid under asymptotics where the proportion of missing data goes to zero as the sample size increases, a form of asymptotics that is not relevant because it does not reflect the loss of information from missing data. In other words, maximizing L_{misp} over $(\theta, \psi, y_{(1)})$ is an example of situation S2 in Section 6.1.3, where the usual asymptotics of ML do not apply.

6.3.2 Background

The classic example of the approach in Section 6.3.1 is the treatment of missing plots in the ANOVA where missing outcomes $y_{(1)}$ are treated as parameters and estimated along with the model parameters to allow computationally efficient methods to be used for analysis (see Chapter 2). DeGroot and Goel (1980) propose this approach as one possibility for the analysis of a mixed-up bivariate normal sample, where the missing data are the indices that allow the values of the two variables to be paired, and *a priori* all pairings are assumed equally likely. Press and Scott (1976) present a Bayesian analysis of an incomplete multivariate normal sample, which is equivalent to maximizing L_{misp} in Eq. (6.68) over $(\theta, y_{(1)})$. Box et al. (1970) and Bard (1974) apply the same approach in a more general setting where the multivariate normal mean vector has a nonlinear regression on covariates. Lee and Nelder (1996) advocate this approach for the analysis of generalized linear mixed models, as discussed in Example 6.30 below.

Formally, $L_{\text{full}}(\theta, \psi \mid y_{(0)}, y_{(1)}, m)$ defined in Eq. (6.47), or $L_{\text{ign}}(\theta \mid y_{(0)})$ defined in Eq. (6.48) if the missingness mechanism is ignorable, define the correct likelihood for inferences about θ based on the observed data; the functions L_{misp} in Eqs. (6.68) or (6.69) are not likelihoods because their arguments include random variables $y_{(1)}$, which have a distribution under the specified model and hence should not be treated as fixed parameters. Thus, maximization of L_{misp} with respect to θ and Y_{mis} is *not* an ML procedure, and as noted in the previous section, it does not generally enjoy the optimal properties of ML. The deficiencies of treating $y_{(1)}$ as a parameter are illustrated in the following examples.

6.3.3 Examples

Example 6.27 Univariate Normal Sample with Missing Data. Suppose that $y = (y_{(0)}, y_{(1)})$ consists of n realizations from a normal distribution with mean μ and variance σ^2 , where $y_{(0)}$ consists of r observed values, and $y_{(1)}$ consists of $(n - r)$ missing values, which are assumed MAR; θ is (μ, σ^2) , which we assume is distinct from the parameters of the missingness mechanism. The ignorable likelihood $L_{\text{ign}}(\theta | y_{(0)})$ is the likelihood for a sample of size r from a normal distribution, and maximizing it over θ yields ML estimates

$$\hat{\mu} = \sum_{i=1}^r \frac{y_i}{r} \quad \text{and} \quad \hat{\sigma}^2 = \sum_{i=1}^r \frac{(y_i - \hat{\mu})^2}{r}. \quad (6.70)$$

Equation (6.68) for this model is

$$\begin{aligned} L_{\text{misp}}(\mu, \sigma^2, y_{r+1}, \dots, y_n | y_1, \dots, y_r) \\ = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^r \frac{(y_i - \mu)^2}{\sigma^2} \right\} \exp \left\{ -\frac{1}{2} \sum_{i=r+1}^n \frac{(y_i - \mu)^2}{\sigma^2} \right\}, \end{aligned} \quad (6.71)$$

and maximizing (6.69) over $\mu, \sigma^2, y_{r+1}, \dots, y_n$ yields estimates $\hat{\mu}^*, \hat{\sigma}^{*2}, \hat{y}_{r+1}, \dots, \hat{y}_n$ where

$$\hat{y}_i = \hat{\mu}, i = r + 1, \dots, n, \quad \hat{\mu}^* = \hat{\mu}, \quad \hat{\sigma}^{*2} = r\hat{\sigma}^2/n, \quad (6.72)$$

where $\hat{\mu}, \hat{\sigma}^2$ are the ML estimates (6.70). Thus, maximizing L_{misp} yields the ML estimate of the mean, but the estimate of the variance is the ML estimate multiplied by the fraction of data observed, r/n . When the fraction of missing data is substantial (e.g., $(n - r)/n = 0.5$), the estimated variance is badly biased, and this bias does not disappear as $n \rightarrow \infty$ unless $r/n \rightarrow 1$; more relevant asymptotics would fix r/n as the sample size increases.

Example 6.28 Missing Plot Analysis of Variance. Suppose we add to the previous example a set of covariates X that is observed for all n observations. We assume that the outcome y_i for observation i with covariate values x_i is normal with mean $\beta_0 + x_i\beta$ and variance σ^2 , and write $\theta = (\beta_0, \beta, \sigma^2)$. The estimates of β_0 , β , and σ^2 that maximize the likelihood $L_{\text{ign}}(\theta | y_{(0)})$ are obtained by applying least squares regression to the data from the r observed units. The estimates of β_0 and β obtained by maximizing L_{misp} are the same as the ML estimates. As in Example 6.26, however, the estimate of variance is the ML estimate multiplied by the proportion of values observed.

Example 6.29 *An Exponential Sample with Censored Values.* In Examples 6.27 and 6.28, estimation based on maximizing L_{mispars} yields reasonable estimates of the location parameters, even though estimates of the scale parameter generally need adjustment. In other examples, however, estimates of location also can be badly biased. For example, consider, as in Example 6.23, a censored sample from an exponential distribution with mean θ , where $y_{(0)}$ represents the r observed values, which lie below a known censoring point c , and $y_{(1)}$ represents the $n - r$ values beyond c , which are censored. As discussed in Example 6.23, the ML estimate of θ is $\hat{\theta} = \bar{y} + (n - r)c/r$. Maximizing L_{mispars} in Eq. (6.68) over θ and $y_{(1)}$ leads to estimating censored values of $\hat{y}_i = c$, $i = r + 1, \dots, n$, and estimating θ by $(r/n)\hat{\theta}$. Thus, in this example, the estimate of the mean is inconsistent unless the proportion of missing values tends to zero as the sample size increases.

Example 6.30 *ML Estimation for Generalized Linear Mixed Models.* Breslow and Clayton (1993) consider an extension of the Generalized Linear Model of Example 6.11 to include random effects. Suppose that conditional on covariates x_i and an unobserved random effect u_i for unit i , the outcome y_i has the distribution of Eq. (6.9), that is

$$f(y_i | x_i, u_i, \beta, \phi) = \exp[(y_i \delta(x_i, u_i, \beta) - b(\delta(x_i, u_i, \beta))) / \phi + c(y_i, \phi)]. \quad (6.73)$$

The mean of y_i given x_i and u_i , $\mu_i = E(y_i | x_i, u_i, \beta, \phi)$, is related to u_i and the covariates x_i by the expression

$$\mu_i = g^{-1} \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} + u_i \right), \quad (6.74)$$

where $g(\cdot)$ is the link function. Also, the u_i are assumed independent with density $f(u_i | x_i, \alpha)$ indexed by unknown parameters α . The likelihood given data $y_{(0)} = (y_1, y_2, \dots, y_n)$ is then:

$$L(\beta, \phi, \alpha | y_{(0)}) = \prod_{i=1}^n \int f(y_i | x_i, u_i, \beta, \phi) f(u_i | x_i, \alpha) du_i, \quad (6.75)$$

which is complicated by the integration over the unobserved random effects u_i , which must be viewed as missing values for likelihood inference. Lee and Nelder (1996) avoid the integration by maximizing

$$h(\beta, \phi, \alpha, u | Y) = \sum_{i=1}^n \log(f(y_i | x_i, u_i, \beta, \phi)) + \log(f(u_i | x_i, \alpha)), \quad (6.76)$$

with respect to (β, ϕ, α) and $u = (u_1, \dots, u_n)^T$. This is the logarithm of L_{mispars} in Eq. (6.68), with $\theta = (\beta, \phi, \alpha)$, $y_{(0)} = (y_1, \dots, y_n)$, and $y_{(1)} = (u_1, \dots, u_n)$. When

both distributions are normal, (6.76) is the joint loglikelihood maximized by Henderson (1975). Unlike maximization of (6.75), maximization of (6.76) does not generally give consistent estimates of the parameters (Breslow and Lin 1995), especially for problems involving nonconjugate distributions for u_i . Algorithms for maximizing the loglikelihood (6.75) are discussed in Breslow and Clayton (1993), McCulloch (1997), and Aitkin (1999).

To address this issue, Lee and Nelder (2001) and Lee et al. (2006) propose maximizing a modification of (6.76), namely a function they call the “adjusted profile h -likelihood”

$$\text{APHL}(\beta, \phi, \alpha) = [h(\beta, \phi, \alpha, u | Y) - 0.5 \log \det \{D(h, u)/2\pi\}]|_{u=\tilde{u}}, \quad (6.77)$$

where $D(h, u) = -\partial^2 h / \partial u^2$ and \tilde{u} solves $\partial h / \partial u = 0$. This approach can be viewed as a numerical (Laplace) approximation to maximizing (6.75), which is the correct ML approach. For more details, see Lee and Nelder (2009) and the discussion, particularly Meng (2009).

6.4 Likelihood Theory for Coarsened Data

Missing values are a form of “data coarsening.” Heitjan and Rubin (1990) and Heitjan (1994) develop a more general theory for coarsened data, which includes heaped, censored, and grouped data as well as missing data. Denoted by $Y = \{y_{ij}\}$, the complete data matrix in the absence of coarsening, and let $f_Y(y | \theta)$ denote the density of Y under a complete data model with unknown parameter θ . The observed data, say $y_{ij(0)}$, for each value y_{ij} are that y_{ij} lies in a subset of its sample space Ψ_{ij} , which is a determined by a function of y_{ij} and a coarsening variable c_{ij} , that is $y_{ij(0)} = y_{ij(0)}(y_{ij}, c_{ij})$, subject to the condition that the coarsened subset contains the unobserved true value, that is $y_{ij} \in y_{ij(0)}(y_{ij}, c_{ij})$. For the special case of missing data discussed so far, $C = \{c_{ij}\}$ is simply the matrix of binary missingness indicators, and

$$y_{ij(0)} = \begin{cases} \{y_{ij}\}, & \text{the set consisting of the single true value, if } c_{ij} = 0, \\ \Psi_{ij}, & \text{the sample space of } y_{ij}, \end{cases} \quad \text{if } c_{ij} = 1.$$

Example 6.31 Censoring with Stochastic Censoring Time. Suppose Y is time to an event, and some values of Y are observed, and others are known to be censored. For observation i , let y_i be the value of Y , and c_i the value of the stochastic censoring time C . The complete data are (y_i, c_i) , $i = 1, \dots, n$. The coarsened data for subject i are

$$y_{(0)i} = y_{(0)i}(y_i, c_i) = \begin{cases} \{y_i\}, & \text{if } y_i \leq c_i; \\ (c_i, \infty), & \text{if } y_i > c_i. \end{cases}$$

That is if the event occurs before the censoring point, then $y_{(0)i}$ is the point set consisting of the actual event time y_i , and if the event occurs after the censoring time, then $y_{(0)i}$ is the set of times beyond c_i . Note that when y_i is observed, the corresponding censoring time c_i is not observed, unlike the missing-data case where the missingness indicator is always observed.

Uncertainty in the degree of coarsening is modeled by assigning C a probability distribution with conditional density given $Y = y$ equal to $f_{C|Y}(c|y, \phi)$. Write $y = (y_{(0)}, y_{(1)})$ and $c = (c_{(0)}, c_{(1)})$, where $y_{(0)}$ and $y_{(1)}$ are the observed and missing components of Y , and $c_{(0)}$ and $c_{(1)}$ are the observed and missing components of C . The full coarsened data likelihood is then

$$L_{\text{full}}(\theta, \phi | y_{(0)}, c_{(0)}) = \iint f_{C|Y}(c_{(0)}, c_{(1)} | y_{(0)}, y_{(1)}, \phi) f_Y(y_{(0)}, y_{(1)} | \theta) dy_{(1)} dc_{(1)}, \quad (6.78)$$

and the likelihood ignoring the coarsening mechanism is

$$L_{\text{ign}}(\theta | y_{(0)}) = \int f_Y(y_{(0)}, y_{(1)} | \theta) dy_{(1)}. \quad (6.79)$$

The following definitions and lemma generalize the ideas of MAR and ignorable missingness mechanisms to coarsened data:

Definition 6.7 The data are coarsened at random (CAR) at the observed values $y_{(0)} = \tilde{y}_{(0)}$, $c_{(0)} = \tilde{c}_{(0)}$ if

$$f_{C|Y}(\tilde{c}_{(0)}, c_{(1)} | \tilde{y}_{(0)}, y_{(1)}, \phi) = f_{C|Y}(\tilde{c}_{(0)}, c_{(1)}^* | \tilde{y}_{(0)}, y_{(1)}^*, \phi) \\ \text{for all } c_{(1)}, c_{(1)}^*, y_{(1)}, y_{(1)}^*, \phi.$$

Definition 6.8 The coarsening mechanism is ignorable if inference for θ based on L_{ign} is equivalent to inference based on the full likelihood L_{full} .

Conditions for ignoring the coarsening mechanism parallel the conditions for ignoring the missingness mechanism, described previously in Theorems 6.1A and 6.1B and Corollaries 6.1A and 6.1B. In particular, sufficient conditions for ignoring the coarsening mechanism for likelihood inference at $\tilde{y}_{(0)}$ and $\tilde{c}_{(0)}$ are that (i) the data are CAR and (ii) the parameters θ and ϕ are distinct. Sufficient conditions for ignoring the coarsening mechanism for Bayesian inference are that (i) the data are CAR and (ii) the parameters θ and ϕ have independent prior distributions.

Example 6.32 Censoring Mechanisms (Example 6.30 Continued). For the case of censored data and (y_i, c_i) independent of (y_j, c_j) when $i \neq j$, the full likelihood equation (6.78) is

$$L_{\text{full}}(\theta, \phi | y_{(0)}, c_{(0)}) = \prod_{i:c_i \geq y_i} \int_{c_i > y_i} f_Y(y_i | x_i, \theta) f_{C|Y}(c_i | y_i, x_i, \phi) dc_i \prod_{i:c_i < y_i} \left(\int_{y_i > c_i} f_{C|Y}(c_i | y_i, x_i, \phi) f_Y(y_i | x_i, \theta) dy_i \right), \quad (6.80)$$

where x_i denotes a set of fully observed covariates for unit i , with $f_Y(y_i | x_i, \theta)$ the density of y_i given x_i and $f_{C|Y}(c_i | y_i, x_i, \phi)$ the density of c_i given (x_i, y_i) . The likelihood equation (6.79) ignoring the coarsening mechanism is

$$L_{\text{ign}}(\theta | y_{(0)}) = \prod_{i:c_i \geq y_i} f_Y(y_i | x_i, \theta) \prod_{i:c_i < y_i} \left(\int_{y_i > c_i} f_Y(y_i | x_i, \theta) dy_i \right). \quad (6.81)$$

The data are CAR at the observed data $\{(\tilde{c}_i, \tilde{y}_i, \tilde{x}_i), i = 1, \dots, n\}$ if

$$f_C(\tilde{c}_i | \tilde{y}_i, \tilde{x}_i, \phi) = f_C(\tilde{c}_i | \tilde{x}_i, \phi) \quad \text{for all } \phi, i = 1, \dots, n$$

because otherwise, in general, the integrals in Eq. (6.80) cannot get passed over the first factors to yield Eq. (6.81). Hence, the censoring mechanism cannot depend on the values of the outcome Y , although it can depend on the values of the covariates. If the distinctness condition is also satisfied, then the censoring mechanism is ignorable for likelihood inference. Note that the censoring mechanism is CAR but not MAR under these conditions. For more discussion, see Heitjan (1994) and Jacobsen and Keiding (1995).

Problems

- 6.1 Write the likelihood function for an iid sample from the (a) beta distribution; (b) Poisson distribution; and (c) Cauchy distribution with location θ and scale 1.
- 6.2 Find the score function for the distributions in Problem 6.1. Find the ML estimates for those distributions that have closed-form estimates.
- 6.3 For a univariate normal sample, find the ML estimate of the coefficient of variation, σ/μ .

- 6.4** (a) Compare ML and least squares estimates for the model of Example 6.10.
 (b) Show that if the data are iid with the Laplace (double exponential) distribution,

$$f(y_i | \theta) = 0.5\exp(-|y_i - \mu(x_i)|),$$

where $\mu(x_i) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$, then ML estimates of β_0, \dots, β_p are obtained by minimizing the sum of absolute deviations of the y values from their expected values.

- 6.5** Suppose the data are a random sample of size n from the uniform distribution between 0 and θ , $\theta > 0$. Show that the ML estimate of θ is the largest data value. (*Hint:* Differentiation of the score function does not work for this problem!) Find the posterior mean of θ , assuming a uniform prior distribution for θ . Which of these estimators do you prefer for this problem and why?
- 6.6** Show that for the GLIM model of Eq. (6.9), $E(y_i | x_i, \beta) = b'(\delta(x_i, \beta))$, where prime denotes differentiation with respect to the function's argument. Conclude that the canonical link Eq. (6.12) is obtained by setting $g^{-1}(\cdot) = b'(\cdot)$. (*Hint:* Consider the density of y_i in Eq. (6.9) as a function of $\delta_i = \delta(x_i, \beta)$ and differentiate the expression $\int f(y_i | \delta_i, \phi) dy_i = 1$ with respect to δ_i ; you may assume that the derivative can be passed through the integral sign.)
- 6.7** Show, by similar arguments to those in Problem 6.6, that for the model of Eq. (6.9), $\text{Var}(y_i | \delta_i, \phi) = \phi b''(\delta_i)$, where $\delta_i = \delta(x_i, \beta)$, and double prime denotes differentiation twice with respect to the function's argument.
- 6.8** Summarize the theoretical and practical differences between the frequentist and Bayesian interpretation of Approximation 6.1. Which is closer to the direct-likelihood interpretation?
- 6.9** For the distributions of Problem 6.1, calculate the observed information and the expected information.
- 6.10** Show that, for random samples from “regular” distributions (differentials can be passed through the integral), the expected squared-score function equals the expected information.
- 6.11** In Example 6.15, show that for large n , $\text{LR} = t^2$.

- 6.12** Derive the posterior distributions in Eqs. (6.26)–(6.28) for Example 6.16.
- 6.13** Derive the posterior distributions in Eqs. (6.34)–(6.36) for Example 6.17.
- 6.14** Derive the modifications of the posterior distributions in Eqs. (6.34)–(6.36) for weighted linear regression, discussed at the end of Example 6.17. Show that for the special case of weighted linear regression with no intercept ($\beta_0 = 0$), a single covariate X , and weight for observation i $w_i = x_i$, the ratio estimator \bar{y}/\bar{x} is (a) the ML estimate of β_1 and (b) the posterior mean of β_1 when the prior distribution for (β_1, σ^2) is $p(\beta_1, \log \sigma^2) = \text{const.}$
- 6.15** Suppose the following data are a random sample of $n = 7$ from the Cauchy distribution with median $\theta : Y = (-4.2, -3.2, -2.0, 0.5, 1.5, 1.5, 3.5)$. Compute and compare 90% intervals for θ using (a) the asymptotic distribution based on the observed information, (b) the asymptotic distribution based on the expected information, and (c) the posterior distribution assuming a uniform prior distribution for θ .
- 6.16** Find large-sample sampling variance estimates for the two ML estimates in Example 6.23.
- 6.17** For a bivariate normal sample $(y_{i1}, y_{i2}), i = 1, \dots, n$ on (Y_1, Y_2) with parameters $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})$, missing values of Y_2 , and m_{i2} the missingness indicator for y_{i2} , state for the following missingness mechanisms (i) whether the data are MAR and (ii) whether the missingness mechanism is ignorable for likelihood-based inference.
- $\Pr(m_{i2} = 1 | y_{i1}, y_{i2}, \theta, \psi) = \exp(\psi_0 + \psi_1 y_{i1}) / \{1 + \exp(\psi_0 + \psi_1 y_{i1})\}$,
 $\psi = (\psi_0, \psi_1)$ distinct from θ .
 - $\Pr(m_{i2} = 1 | y_{i1}, y_{i2}, \theta, \psi) = \exp(\psi_0 + \psi_1 y_{i2}) / \{1 + \exp(\psi_0 + \psi_1 y_{i2})\}$,
 $\psi = (\psi_0, \psi_1)$ distinct from θ .
 - $\Pr(m_{i2} = 1 | y_{i1}, y_{i2}, \theta, \psi) = 0.5 \exp(\mu_1 + \psi y_{i1}) / \{1 + \exp(\mu_1 + \psi y_{i1})\}$,
scalar ψ distinct from θ .
- 6.18** Suppose that, given sets of covariates X_1 and X_2 (possibly overlapping, that is, not disjoint), y_{i1} and y_{i2} are bivariate normal with means $x_{i1}\beta_1$ and $x_{i2}\beta_2$, variances σ_1^2 and $\sigma_2^2 = 1$, and correlation ρ . The data consist of a random sample of units i with x_{i1} and x_{i2} always observed, y_{i2} always missing and y_{i1} missing if and only if $y_{i2} > 0$. Let m_{i1} be the missingness indicator for y_{i1} . Show that

$$\Pr(m_{i1} = 1 | y_{i1}, x_{i1}, x_{i2}, \theta, \psi) = 1 - \Phi\left(\frac{-x_{i2}\beta_2 - (\rho/\sigma_1)(y_{i1} - x_{i1}\beta_1)}{\sqrt{1 - \rho^2}}\right),$$

where Φ is the standard normal cumulative distribution function. Hence, give conditions on the parameters under which the data are MAR and under which the missingness mechanism is ignorable for likelihood-based inference. (This model is considered in detail in Example 15.5.)

- 6.19** The definition of MAR can depend on how the complete data are defined. Suppose that $X = (x_1, \dots, x_n)$, $Z = (z_1, \dots, z_n)$ (x_i, z_i) are a random sample from a bivariate normal distribution with means $(\mu_x, 0)$, variances $(\sigma_x^2, 1)$, and correlation 0 (so x_i and z_i are independent). Suppose that some values x_i are missing, the missingness indicator for x_i is m_i , Z are completely unobserved latent variables, and the missingness mechanism is given by

$$\Pr(m_i = 1 | x_i, z_i) = \exp(z_i)/(1 + \exp(z_i)).$$

Show that if the complete data are defined as X , then the missing data are missing completely at random (MCAR), but if the complete data are defined as (X, Z) , then the missing data are not MAR. Which is the more sensible definition?

- 6.20** For Example 6.27, derive the estimates maximizing over parameters and missing data in Eq. (6.71).
- 6.21** For Example 6.28, derive the estimates maximizing over parameters and missing data, as described in the example.
- 6.22** In Example 6.29, derive the estimates maximizing over parameters and missing data.

Notes

- 1 A technically more precise formulation is to write $A^{-1}(\theta - \hat{\theta}) \sim N(0, I_d)$, where A is the matrix square root of C (that is, $A^T A = A A^T = C$) and I_d is the $(d \times d)$ identity matrix.
- 2 More formally, $y_{(0)}$ is an $(n \times K)$ matrix with entries y_{ij} when $m_{ij} = 0$ and $*$ when $m_{ij} = 1$. In earlier editions, we used the notation y_{obs} and y_{mis} for $y_{(0)}$ and $y_{(1)}$, respectively. This notation has led to some confusion, as discussed in Seaman et al. (2013) and Mealli and Rubin (2015), and here, we revert a notation closer to the original notation in Rubin (1976a).
- 3 In earlier work (Rubin 1976a; Little and Rubin 1987), the term “distinctness” was used in the Bayesian setting to refer to *a priori* independence, but here, we simply label the condition as “*a priori* independence.”

7

Factored Likelihood Methods When the Missingness Mechanism Is Ignorable

7.1 Introduction

We now assume that the missingness mechanism is ignorable, and for simplicity, write $\ell(\theta | y_{(0)})$ for the ignorable loglikelihood $\ell_{\text{ign}}(\theta | y_{(0)})$ based on the observed data $y_{(0)}$. This can be a complicated function with no obvious maximum and an apparently complex form for the information matrix. Analogous issues arise with simulating from the resulting posterior distributions without simple structures. For certain models and incomplete data patterns, however, analyses based on $\ell(\theta | y_{(0)})$ can employ standard complete-data techniques. The general idea will be described here in this section, and then, specific examples will be given in the remainder of this chapter for normal data and in Section 12.2 for multinomial (that is, cross-classified) data.

For a variety of models and missingness patterns, an alternative parameterization $\phi = \phi(\theta)$, where ϕ is a one-to-one function of θ , can be found such that the loglikelihood decomposes into J terms

$$\ell(\phi | y_{(0)}) = \ell_1(\phi_1 | y_{(0)}) + \ell_2(\phi_2 | y_{(0)}) + \cdots + \ell_J(\phi_J | y_{(0)}), \quad (7.1)$$

where

- $\phi_1, \phi_2, \dots, \phi_J$ are distinct parameters, in the sense that the joint parameter space of $\phi = (\phi_1, \phi_2, \dots, \phi_J)$ is the product of the individual parameter spaces for $\phi_j, j = 1, \dots, J$; and
- each component $\ell_j(\phi_j | y_{(0)})$ corresponds to a loglikelihood for a complete-data problem, or more generally, for an incomplete-data problem that is easier to analyze than that based on $\ell(\theta | y_{(0)})$.

For a Bayesian analysis with a prior distribution specified for the parameters, the aforementioned condition (1) is replaced by the requirement that $\phi_1, \phi_2, \dots, \phi_J$ are mutually *a priori* independent.

If a factored likelihood with these properties can be found, then $\ell(\phi | y_{(0)})$ can be maximized by maximizing $\ell_j(\phi_j | y_{(0)})$ separately for each j . If $\hat{\phi}$ is the resulting maximum likelihood (ML) estimate of ϕ , then the ML estimate of any function $\theta(\phi)$ of ϕ is obtained by applying Property 6.1, that is, substituting $\hat{\theta} = \theta(\hat{\phi})$. Similarly, for Bayesian inference, when (1) and (2) hold, the posterior distribution of ϕ is the product of J independent posterior distributions of $\phi_1, \phi_2, \dots, \phi_J$, and hence, often has a much simpler form than the posterior distribution of θ . The posterior distribution of θ can be simulated by generating draws $\{\phi^{(d)}, d = 1, \dots, D\}$ from the posterior distribution of ϕ and then computing $\{\theta^{(d)} = \theta(\phi^{(d)}), d = 1, \dots, D\}$, which is a set of draws from the posterior distribution of θ by Property 6.1B.

The decomposition (7.1) can also be used to calculate the large-sample covariance matrix associated with the ML estimates, as given in Approximations 6.1 and 6.2. Differentiating (7.1) twice with respect to ϕ_1, \dots, ϕ_J yields a block-diagonal information matrix for ϕ of the form

$$I(\phi | y_{(0)}) = \begin{bmatrix} I(\phi_1 | y_{(0)}) & & & 0 \\ & I(\phi_2 | y_{(0)}) & & \\ & & \ddots & \\ 0 & & & I(\phi_J | y_{(0)}) \end{bmatrix}.$$

Hence, the large sample covariance matrix for $\phi - \hat{\phi}$ is also block diagonal, with the form

$$C(\phi - \hat{\phi} | y_{(0)}) = \begin{bmatrix} I^{-1}(\hat{\phi}_1 | y_{(0)}) & & & 0 \\ & I^{-1}(\hat{\phi}_2 | y_{(0)}) & & \\ & & \ddots & \\ 0 & & & I^{-1}(\hat{\phi}_J | y_{(0)}) \end{bmatrix}. \quad (7.2)$$

Because each nonzero component of this matrix corresponds to a complete-data problem, each is often relatively easy to calculate. By Property 6.2, the approximate large-sample covariance matrix of the ML estimate of a function $\theta = \theta(\phi)$ of ϕ can be found using the formula

$$C(\theta - \hat{\theta} | \tilde{y}_{(0)}) = D(\hat{\theta}) C(\hat{\phi} | \tilde{y}_{(0)}) D^T(\hat{\theta}), \quad (7.3)$$

where $D(\cdot)$ is the matrix of partial derivatives of the column vector θ with respect to ϕ :

$$D(\theta) = \{d_{jk}(\theta)\}, \quad \text{where } d_{jk}(\theta) = \frac{\partial \theta_j}{\partial \phi_k}.$$

7.2 Bivariate Normal Data with One Variable Subject to Missingness: ML Estimation

7.2.1 ML Estimates

Anderson (1957) first introduced factored likelihoods for the normal data of Example 6.24.

Example 7.1 *Bivariate Normal Sample with One Variable Subject to Nonresponse (Example 6.24 Continued).* The loglikelihood for a bivariate normal sample with r complete bivariate units $\{(y_{i1}, y_{i2}), i = 1, \dots, r\}$ and $n - r$ univariate units $\{y_{i1}, i = r + 1, \dots, n\}$ is given by (6.52). ML estimates of μ and Σ are found by maximizing this function with respect to μ and Σ . The likelihood equations, however, do not have an obvious solution. Anderson (1957) factors the joint distribution of y_{i1} and y_{i2} into the marginal distribution of y_{i1} and the conditional distribution of y_{i2} given y_{i1} :

$$f_Y(y_{i1}, y_{i2} | \mu, \Sigma) = f_1(y_{i1} | \mu_1, \sigma_{11}) f_2(y_{i2} | y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1}),$$

where, by properties of the bivariate normal distribution discussed in Example 6.9, $f_1(y_{i1} | \mu_1, \sigma_{11})$ is the univariate normal distribution with mean μ_1 and variance σ_{11} , and $f_2(y_{i2} | y_{i1}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})$ is the univariate conditional normal distribution with mean

$$\beta_{20.1} + \beta_{21.1} y_{i1}$$

and variance $\sigma_{22.1}$. The parameter

$$\phi = (\mu_1, \sigma_{11}, \beta_{20.1}, \beta_{21.1}, \sigma_{22.1})^T$$

is a one-to-one function of the original parameter

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$$

of the joint distribution of y_{i1} and y_{i2} . In particular, μ_1 and σ_{11} are common to both parameterizations, and the other components of ϕ are given by the following functions of components of θ :

$$\begin{aligned} \beta_{21.1} &= \sigma_{12}/\sigma_{11}, \\ \beta_{20.1} &= \mu_2 - \beta_{21.1}\mu_1, \\ \sigma_{22.1} &= \sigma_{22} - \sigma_{12}^2/\sigma_{11}. \end{aligned} \tag{7.4}$$

Similarly, the components of θ other than μ_1 and σ_{11} can be expressed as the following functions of the components of ϕ :

$$\begin{aligned}\mu_2 &= \beta_{20\cdot 1} + \beta_{21\cdot 1}\mu_1, \\ \sigma_{12} &= \beta_{21\cdot 1}\sigma_{11}, \\ \sigma_{22} &= \sigma_{22\cdot 1} + \beta_{21\cdot 1}^2\sigma_{11}.\end{aligned}\tag{7.5}$$

The density of the data $y_{(0)}$ factors in the following way:

$$\begin{aligned}f(y_{(0)} | \theta) &= \prod_{i=1}^r f_Y(y_{i1}, y_{i2} | \theta) \prod_{i=r+1}^n f_1(y_{i1} | \theta) \\ &= \left[\prod_{i=1}^r f_1(y_{i1} | \theta) f_2(y_{i2} | y_{i1}, \theta) \right] \left[\prod_{i=r+1}^n f_1(y_{i1} | \theta) \right] \\ &= \left[\prod_{i=1}^n f_1(y_{i1} | \mu_1, \sigma_{11}) \right] \left[\prod_{i=1}^r f_2(y_{i2} | y_{i1}, \beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1}) \right].\end{aligned}\tag{7.6}$$

The first bracketed factor in the last row of (7.6) is the density of an independent and identically distributed sample of size n from the normal distribution with mean μ_1 and variance σ_{11} . The second factor is the density for r independent units from the conditional normal distribution with mean $\beta_{20\cdot 1} + \beta_{21\cdot 1}y_{i1}$ and variance $\sigma_{22\cdot 1}$. Furthermore, if the parameter space for θ is the standard natural parameter space with no prior restrictions, then (μ_1, σ_{11}) and $(\beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1})$ are distinct, implying that knowledge about (μ_1, σ_{11}) does not provide any knowledge about $(\beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1})$. Hence, ML estimates of ϕ can be obtained by independently maximizing the likelihoods corresponding to these two sets of components.

Maximizing the first factor yields

$$\hat{\mu}_1 = n^{-1} \sum_{i=1}^n y_{i1}, \quad \hat{\sigma}_{11} = \sum_{i=1}^n (y_{i1} - \hat{\mu}_1)^2 / n.\tag{7.7}$$

Maximizing the second factor uses standard regression results and yields estimates from the r complete units (cf. Example 6.9):

$$\begin{aligned}\hat{\beta}_{21\cdot 1} &= \tilde{\sigma}_{12}/\tilde{\sigma}_{11}, \\ \hat{\beta}_{20\cdot 1} &= \bar{y}_2 - \hat{\beta}_{21\cdot 1}\bar{y}_1, \\ \hat{\sigma}_{22\cdot 1} &= \tilde{\sigma}_{22},\end{aligned}\tag{7.8}$$

where $\bar{y}_j = r^{-1} \sum_{i=1}^r y_{ij}$, $\tilde{\sigma}_{jk} = \sum_{i=1}^r (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)/r$ for $j, k = 1, 2$ and $\tilde{\sigma}_{22\cdot 1} = \tilde{\sigma}_{22} - \tilde{\sigma}_{12}^2/\tilde{\sigma}_{11}$ are the ML estimates of their respective parameters if only the first r units were observed. The ML estimates of other parameters can now be obtained using Property 6.1. In particular, from (7.5),

$$\hat{\mu}_2 = \hat{\beta}_{20\cdot 1} + \hat{\beta}_{21\cdot 1}\hat{\mu}_1,$$

or from (7.7) and (7.8),

$$\hat{\mu}_2 = \bar{y}_2 + \hat{\beta}_{21 \cdot 1}(\hat{\mu}_1 - \bar{y}_1); \quad (7.9)$$

from (7.5),

$$\hat{\sigma}_{22} = \tilde{\sigma}_{22 \cdot 1} + \hat{\beta}_{21 \cdot 1}^2 \hat{\sigma}_{11},$$

or from (7.7) and (7.8),

$$\hat{\sigma}_{22} = \tilde{\sigma}_{22} + \hat{\beta}_{21 \cdot 1}^2 (\hat{\sigma}_{11} - \tilde{\sigma}_{11}). \quad (7.10)$$

Finally, from (7.5), we have for the correlation

$$\rho \equiv \sigma_{12}(\sigma_{11}\sigma_{22})^{-1/2} = \beta_{21 \cdot 1}\sigma_{11}^{1/2} (\sigma_{22 \cdot 1} + \hat{\beta}_{21 \cdot 1}^2\sigma_{11})^{-1/2},$$

so from (7.7) and (7.8),

$$\hat{\rho} = \tilde{\rho}(\hat{\sigma}_{11}/\tilde{\sigma}_{11})^{1/2}(\tilde{\sigma}_{22}/\hat{\sigma}_{22})^{1/2}, \quad (7.11)$$

where $\tilde{\rho} = \tilde{\sigma}_{12}(\tilde{\sigma}_{11}\tilde{\sigma}_{22})^{-1/2}$. The first terms on the right side of (7.9) and (7.10) and the first factor on the right side of (7.11) are the ML estimates of μ_2 , σ_{22} , and ρ with the $n - r$ incomplete units discarded. The remaining terms and factors in these expressions represent adjustments based on the additional information from the $n - r$ extra values of y_{i1} .

The ML estimate (7.9) of the mean of y_{i2} is of particular interest. It can be written in the form

$$\hat{\mu}_2 = n^{-1} \left\{ \sum_{i=1}^r y_{i2} + \sum_{i=r+1}^n \hat{y}_{i2} \right\}, \quad (7.12)$$

where

$$\hat{y}_{i2} = \bar{y}_2 + \hat{\beta}_{21 \cdot 1}(y_{i1} - \bar{y}_1).$$

Hence, $\hat{\mu}_2$ is a type of regression estimator commonly used in sample surveys (e.g., Cochran 1977), which effectively imputes the predicted values \hat{y}_{i2} for missing y_{i2} from the linear regression of observed y_{i2} on observed y_{i1} .

Example 7.2 Bivariate Normal Numerical Illustration. The first $r = 12$ units in Table 7.1, taken from Snedecor and Cochran (1967, table 6.9.1), give measurements on the size of crop from apple trees, in hundreds of fruits (y_{i1}) and 100 times the percentage of wormy fruits (y_{i2}). These units suggest a negative association between the size of crop and the percentage of wormy fruits. Suppose that the objective is to estimate the mean of y_{i2} , but the value of y_{i2} was not determined for some of the trees with smaller crops, numbered 13–18 in the table. The sample mean for the complete units, $\bar{y}_2 = 45$, underestimates the

Table 7.1 Example 7.2: data on size of apple crop (y_{i1}) and 100 × percentage of wormy fruit (y_{i2})

Tree number	Size of crop (100s of fruits) (y_{i1})	100 × percentage wormy fruits (y_{i2})	Regression prediction (y_{i2})
1	8	59	56.1
2	6	58	58.2
3	11	56	53.1
4	22	53	42.0
5	14	50	50.1
6	17	45	47.0
7	18	43	46.0
8	24	42	39.9
9	19	39	45.0
10	23	38	41.0
11	26	30	37.9
12	40	27	23.7
13	4	?	60.2
14	4	?	60.2
15	5	?	59.2
16	6	?	58.2
17	8	?	56.1
18	10	?	54.1

$$\bar{y}_1 = 19; \quad \bar{y}_2 = 45; \quad \hat{\mu}_2 = 49.3333; \quad \hat{\mu}_1 = 14.7222$$

$$\tilde{\sigma}_{11} = 77.0; \quad \tilde{\sigma}_{12} = -78.0; \quad \tilde{\sigma}_{22} = 101.8333; \quad \hat{\sigma}_{11} = 89.5340$$

? denotes missing.

Source: Snedecor and Cochran (1967, table 6.9.1). Adapted with permission of Iowa State University Press.

percentage of wormy fruits because the percentage for the six omitted trees, which tend to be smaller, is expected to be larger than the percentage for the measured trees (that is the data may be missing at random [MAR] but do not appear to be missing completely at random [MCAR]). The ML estimate assuming ignorability (that is MAR and distinctness of the data and missingness parameters) is $\hat{\mu}_2 = 49.33$, which can be compared with the estimate $\bar{y}_2 = 45$ from the complete units. This analysis should be taken only as a numerical illustration; a serious analysis of these data would consider issues such as whether transformations of y_{i1} and y_{i2} (e.g., log, square root) would better meet the underlying linearity and normality assumptions.

7.2.2 Large-Sample Covariance Matrix

The large-sample covariance matrix of $(\phi - \hat{\phi})$ is found by calculating and inverting the information matrix. The loglikelihood of ϕ is, from Eq. (7.6),

$$\ell(\phi | y_{(0)}) = -(2\sigma_{22 \cdot 1})^{-1} \sum_{i=1}^r (y_{i2} - \beta_{20 \cdot 1} - \beta_{21 \cdot 1} y_{i1})^2 - \frac{1}{2} r \ln \sigma_{22 \cdot 1}$$

$$-(2\sigma_{11})^{-1} \sum_{i=1}^n (y_{i1} - \mu_1)^2 - \frac{1}{2} n \ln \sigma_{11}.$$

Differentiating twice with respect to ϕ gives

$$I(\hat{\phi} | y_{(0)}) = \begin{bmatrix} I_{11}(\hat{\mu}_1, \hat{\sigma}_{11} | y_{(0)}) & 0 \\ 0 & I_{22}(\hat{\beta}_{20 \cdot 1}, \hat{\beta}_{21 \cdot 1}, \hat{\sigma}_{22 \cdot 1} | y_{(0)}) \end{bmatrix},$$

where

$$I_{11}(\hat{\mu}_1, \hat{\sigma}_{11} | y_{(0)}) = \begin{bmatrix} n/\hat{\sigma}_{11} & 0 \\ 0 & n/(2\hat{\sigma}_{11}^2) \end{bmatrix},$$

and

$$I_{22}(\hat{\beta}_{20 \cdot 1}, \hat{\beta}_{21 \cdot 1}, \hat{\sigma}_{22 \cdot 1} | y_{(0)}) = \begin{bmatrix} r/\hat{\sigma}_{22 \cdot 1} & r\bar{y}_1/\hat{\sigma}_{22 \cdot 1} & 0 \\ r\bar{y}_1/\hat{\sigma}_{22 \cdot 1} & \sum_{i=1}^r y_{i1}^2/\hat{\sigma}_{22 \cdot 1} & 0 \\ 0 & 0 & r/(2\hat{\sigma}_{22 \cdot 1}^2) \end{bmatrix}.$$

Inverting these matrices yields the large-sample covariance matrix of $(\phi - \hat{\phi})$ as follows:

$$C(\phi - \hat{\phi}) = \begin{bmatrix} I_{11}^{-1}(\hat{\mu}_1, \hat{\sigma}_{11} | y_{(0)}) & 0 \\ 0 & I_{22}^{-1}(\hat{\beta}_{20 \cdot 1}, \hat{\beta}_{21 \cdot 1}, \hat{\sigma}_{22 \cdot 1} | y_{(0)}) \end{bmatrix},$$

where

$$I_{11}^{-1}(\hat{\mu}_1, \hat{\sigma}_{11} | y_{(0)}) = \begin{bmatrix} \hat{\sigma}_{11}/n & 0 \\ 0 & 2\hat{\sigma}_{11}^2/n \end{bmatrix}$$

and

$$I_{22}^{-1}(\hat{\beta}_{20 \cdot 1}, \hat{\beta}_{21 \cdot 1}, \hat{\sigma}_{22 \cdot 1} | y_{(0)}) = \begin{bmatrix} \hat{\sigma}_{22 \cdot 1}(1 + \bar{y}_1^2/\tilde{\sigma}_{11})/r & -\bar{y}_1\hat{\sigma}_{22 \cdot 1}/(r\tilde{\sigma}_{11}) & 0 \\ -\bar{y}_1\hat{\sigma}_{22 \cdot 1}/(r\tilde{\sigma}_{11}) & \hat{\sigma}_{22 \cdot 1}/(rs\tilde{\sigma}_{11}) & 0 \\ 0 & 0 & 2\hat{\sigma}_{22 \cdot 1}^2/r \end{bmatrix}.$$

The large sample covariance matrix of $(\theta - \hat{\theta})$ can be found using Eq. (7.3). To illustrate the calculations, we consider the parameter μ_2 , the mean of the incompletely observed variable. Because $\mu_2 = \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1$, we have

$$\begin{aligned} D(\mu_2) &= \left(\frac{\partial \mu_2}{\partial \mu_1}, \frac{\partial \mu_2}{\partial \sigma_{11}}, \frac{\partial \mu_2}{\partial \beta_{20\cdot1}}, \frac{\partial \mu_2}{\partial \beta_{21\cdot1}}, \frac{\partial \mu_2}{\partial \sigma_{22\cdot1}} \right) \\ &= (\hat{\beta}_{21\cdot1}, 0, 1, \hat{\mu}_1, 0), \end{aligned}$$

substituting ML estimates of μ_1 and $\beta_{21\cdot1}$. Hence, with some calculation, the large-sample variance of $(\mu_2 - \hat{\mu}_2)$ is

$$\text{Var}(\mu_2 - \hat{\mu}_2) = D(\hat{\mu}_2)C(\phi - \hat{\phi})D(\hat{\mu}_2)^T = \hat{\sigma}_{22\cdot1} \left[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1 - \hat{\rho}^2)} + \frac{(\bar{y}_1 - \hat{\mu}_1)^2}{r\tilde{\sigma}_{11}} \right]. \quad (7.13)$$

The third term in the brackets is of order $(1/r^2)$ if the data are MCAR because in that case, $(\bar{y}_1 - \hat{\mu}_1)^2$ is of order $(1/r)$. Ignoring this term yields

$$\text{Var}(\mu_2 - \hat{\mu}_2) \approx \hat{\sigma}_{22\cdot1} \left[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1 - \hat{\rho}^2)} \right] = \frac{\hat{\sigma}_{22}}{r} \left(1 - \hat{\rho}^2 \frac{n-r}{n} \right), \quad (7.14)$$

which can be compared with the sampling variance of \bar{y}_2 , namely σ_{22}/r . Thus, in large samples, under MCAR, the proportionate reduction in sampling variance obtained by including the $n-r$ units with only y_1 observed is $\hat{\rho}^2$ times the fraction of incomplete units $(n-r)/n$.

7.3 Bivariate Normal Monotone Data: Small-Sample Inference

Given large samples, interval estimates for the parameters can be obtained by applying Approximation 6.1 (Eq. (6.18)), as discussed in Section 6.1.3. In particular, an asymptotic 95% interval for μ_2 takes the form

$$\hat{\mu}_2 \pm 1.96 \sqrt{\text{Var}(\hat{\mu}_2 - \mu_2)}, \quad (7.15)$$

where $\text{Var}(\hat{\mu}_2 - \mu_2)$ is approximated by Eq. (7.13). For parameters other than means or regression coefficients, better intervals are generally obtained by applying a transformation to approximate normality, calculating a normal-based interval for the transformed parameter, and then transforming the interval back to the original scale (see Property 6.2 and Approximation 6.2 in Section 6.1.3). For example, an appropriate transformation for a variance

parameter is the logarithm, so to compute a 95% interval for σ_{22} , a 95% interval for $\ell n(\sigma_{22})$ is computed as

$$\ell n(\hat{\sigma}_{22}) \pm 1.96 \sqrt{\text{Var}(\ell n(\hat{\sigma}_{22}) - \ell n(\sigma_{22}))}, \quad (7.16)$$

where in large samples, $\text{Var}(\ell n(\hat{\sigma}_{22}) - \ell n(\sigma_{22})) = \text{Var}(\hat{\sigma}_{22} - \sigma_{22})/\hat{\sigma}_{22}^2$. The 95% interval for $\hat{\sigma}_{22}$ is then $(\exp(l), \exp(u))$ where (l, u) is the interval for $\ell n(\hat{\sigma}_{22})$ computed from (7.16).

Small-sample inference is problematic from a frequentist perspective. In particular, the quantity $(\hat{\mu}_2 - \mu_2)/\sqrt{\text{Var}(\hat{\mu}_2 - \mu_2)}$ obtained from (7.13) is standard normally distributed in large samples, but its distribution in small samples is complex and depends on the parameters. The t distribution with $r - 1$ degrees of freedom has been suggested as a useful approximate reference distribution for this quantity and performs reasonably well in simulations (Little 1976). The same reference t distribution has also been proposed for inference about the difference in means $\mu_2 - \mu_1$, based on $(\hat{\mu}_2 - \hat{\mu}_1 - \mu_2 + \mu_1)/\sqrt{\text{Var}(\hat{\mu}_2 - \hat{\mu}_1 - \mu_2 + \mu_1)}$.

A more direct (and in our view, more principled) approach to small-sample interval estimation is to specify a prior distribution for the parameters and then derive the associated posterior distribution. Specifically, suppose $\mu_1, \sigma_{11}, \beta_{20\cdot 1}, \beta_{21\cdot 1}$, and $\sigma_{22\cdot 1}$ are assumed *a priori* independent with convenience prior

$$f(\mu_1, \sigma_{11}, \beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1}) \propto \sigma_{11}^{-a} \sigma_{22\cdot 1}^{-c}. \quad (7.17)$$

The choice $a = c = 1$ yields the Jeffreys' prior for the factored density (Box and Tiao 1973).

Applying standard Bayesian theory to the random sample $\{y_{i1}: i = 1, \dots, n\}$, we have the following results: (i) the posterior distribution of (μ_1, σ_{11}) is such that $n\hat{\sigma}_{11}/\sigma_{11}$ has a chi-squared distribution with $n + 2a - 3$ degrees of freedom, and (ii) the posterior distribution of μ_1 given σ_{11} is normal with mean $\hat{\mu}_1$ and variance σ_{11}/n ; applying standard Bayesian regression theory to the random sample $\{(y_{i1}, y_{i2}): i = 1, \dots, r\}$, (iii) the posterior distribution of $(\beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1})$ is such that $r\hat{\sigma}_{22\cdot 1}/\sigma_{22\cdot 1}$ has a chi-squared distribution with $r + 2c - 4$ degrees of freedom, (iv) the posterior distribution of $\beta_{21\cdot 1}$ given $\sigma_{22\cdot 1}$ is normal with mean $\hat{\beta}_{21\cdot 1}$ and variance $\sigma_{22\cdot 1}/(r\tilde{\sigma}_{11})$, and (v) the posterior distribution of $\beta_{20\cdot 1}$ given $\beta_{21\cdot 1}$ and $\sigma_{22\cdot 1}$ is normal with mean $\bar{y}_2 - \beta_{21\cdot 1}\bar{y}_1$ and variance $\sigma_{22\cdot 1}/r$; furthermore, (vi) (μ_1, σ_{11}) and $(\beta_{20\cdot 1}, \beta_{21\cdot 1}, \sigma_{22\cdot 1})$ are a posteriori independent. For derivations of these results, see for example Lindley (1965) or Gelman et al. (2013).

Results 1–6 and Property 6.1B imply that the posterior distribution of any function of the parameters ϕ can be simulated by creating D draws, $d = 1, \dots, D$, as follows:

1. Draw independently x_{1t}^2 and x_{2t}^2 from chi-squared distributions with $n+2a-3$ and $r+2c-4$ degrees of freedom, respectively. Draw three independent standard normal deviates z_{1t} , z_{2t} , and z_{3t} .
2. Compute $\phi^{(d)} = (\sigma_{11}^{(d)}, \mu_1^{(d)}, \sigma_{22-1}^{(d)}, \beta_{20-1}^{(d)}, \beta_{21-1}^{(d)})^\top$, where

$$\begin{aligned}\sigma_{11}^{(d)} &= n\hat{\sigma}_{11}/x_{1t}^2, \\ \mu_1^{(d)} &= \hat{\mu}_1 + z_{1t}(\sigma_{11}^{(d)}/n)^{1/2}, \\ \sigma_{22-1}^{(d)} &= r\hat{\sigma}_{22-1}/x_{2t}^2, \\ \beta_{21-1}^{(d)} &= \hat{\beta}_{21-1} + z_{2t}(\sigma_{22-1}^{(d)}/(r\tilde{\sigma}_{11}))^{1/2}, \\ \beta_{20-1}^{(d)} &= \bar{y}_2 - \beta_{21-1}^{(d)}\bar{y}_1 + z_{3t}(\sigma_{22-1}^{(d)}/r)^{1/2}.\end{aligned}$$

3. Compute the corresponding transformation of $\phi^{(d)}$. For example, if the transformation is $\mu_2 = \beta_{20-1} + \beta_{21-1}\mu_1$, then $\mu_2^{(d)} = \beta_{20-1}^{(d)} + \beta_{21-1}^{(d)}\mu_1^{(d)}$.

The methods of the previous two sections are now applied to the data in Table 7.1.

Example 7.3 Bayes Interval Estimation for the Bivariate Normal (Example 7.2 Continued). Table 7.2 shows 95% intervals for μ_2 , σ_{22} , and ρ for the data in Table 7.1. Intervals based on four methods are presented:

- (1) Asymptotic intervals based on the inverse of the observed information matrix of $(\mu_2, \sigma_{22}, \rho)$ (e.g., Eq. (7.15) for μ_2);
- (2) Intervals with standard errors based on the inverse of the observed information matrix of $(\mu_2, \ln \sigma_{22}, Z_\rho)$, where $Z_\rho = \ln[(1+\rho)/(1-\rho)]/2$ is Fisher's

Table 7.2 Example 7.3: 95% intervals for parameters of bivariate normal distribution, based on data in Table 7.1

Method	Parameters		
	μ_2	σ_{22}	ρ
(1) Asymptotic theory	(44.0, 54.7)	(30.7, 198.7)	(-1.00, -0.79)
(2) Asymptotic theory, with transform and t approximation	(43.4, 55.3)	(50.8, 258.9)	(-0.97, -0.69)
(3) Bayes simulation A	(43.7, 54.4)	(60.1, 289.7)	(-0.96, -0.66)
(3) Bayes simulation B	(43.5, 55.7)	(59.2, 293.9)	(-0.96, -0.66)
(4) Normal approximation to A	(43.5, 55.2)	(15.9, 256.7)	(-1.03, -0.72)
(4) Normal approximation to B	(43.4, 55.2)	(14.6, 257.2)	(-1.03, -0.71)
ML estimates	49.33	114.70	-0.895

normalizing transformation of the correlation. The normal percentile 1.96 of the asymptotic normal interval is replaced by the 97.5th percentile (2.201) of the t distribution on $r - 1 = 11$ degrees of freedom.

- (3) The 2.5th to 97.5th percentile of the Bayesian posterior distribution with prior distribution (7.17) with $a = c = 1$, simulated using the method of Section 7.3 with 9999 simulated values; and
- (4) Intervals obtained by fitting normal distributions to the simulated posterior distributions from method (3), using the mean and variance of the 9999 simulated values.

Methods 3 and 4 are repeated for two independent sets of random numbers (A and B) to give some idea of the simulation variance.

The asymptotic interval for μ_2 from method (1) is shorter than for the other methods, and presumably has lower than the stated 95% coverage, because uncertainty due to estimation of the variance parameters is not taken into account. The other intervals for μ_2 are fairly similar to each other. The intervals for σ_{22} and ρ that rely on normality on the original scale (methods (1) and (4)) are not satisfactory – in particular, the lower limits of the intervals for the correlation lie outside the parameter space. The intervals for methods (2) and (3) are broadly similar. Method (2) forces symmetry around the ML estimates of $\ln\sigma_{22}$ and Z_ρ , and method (4) forces symmetry about the sample mean of the posterior draws of σ_{22} and ρ : the normal approximations of method (4) should be applied to the draws on the transformed scale and then the interval transformed back to the original scale. Method (3) has the advantage of not imposing these symmetries on the intervals but suffers from simulation error from the finite number of posterior draws, $D = 9999$; this is a minor issue and easily corrected by increasing D . Coverage properties of these intervals in repeated sampling deserve more extensive simulation; for example, Little (1988a) studies coverage properties of various t approximations to the posterior distribution of μ_2 , but current computing environments make more extensive evaluations straightforward.

7.4 Monotone Missingness with More Than Two Variables

7.4.1 Multivariate Data with One Normal Variable Subject to Missingness

A simple but important extension of bivariate monotone missing data is given in the next example.

Example 7.4 *K + 1 Variables, One Subject to Missingness.* Suppose we replace y_{i1} by a set of K completely observed variables, as for the data pattern in Figure 1.1a, resulting in a special case of monotone data with $J = 2$ and Y_1 representing K variables. Suppose first that (y_{i1}, y_{i2}) are iid $(K + 1)$ -variate normally

distributed, and the data are MAR with distinct parameters. The ML estimates of μ_2 and σ_{22} are then

$$\hat{\mu}_2 = \bar{y}_2 + (\hat{\mu}_1 - \bar{y}_1)^T \hat{\beta}_{21.1}, \quad (7.18)$$

and

$$\hat{\sigma}_{22} = s_{22} + \hat{\beta}_{21.1}^T (\hat{\sigma}_{11} - \tilde{\sigma}_{11}) \hat{\beta}_{21.1},$$

where $\hat{\mu}_1$ and \bar{y}_1 are $(K \times 1)$ vectors, $\hat{\mu}_1$ the mean of y_{i1} for all n units, \bar{y}_1 the sample mean of y_{i1} for the r complete units, respectively, $\hat{\beta}_{21.1}$ is the $(K \times 1)$ vector of estimated regression coefficients from the multiple regression of y_{i2} on y_{i1} , and $\hat{\sigma}_{11}$ and $\tilde{\sigma}_{11}$ are the $(K \times K)$ sample covariance matrices, $\hat{\sigma}_{11}$ based on all n units of y_{i1} and $\tilde{\sigma}_{11}$ based on the r units of y_{i1} where y_{i2} is also observed. The ML estimate $\hat{\mu}_2$ corresponds to imputing the missing values of y_{i2} using the ML estimates of the parameters of the multiple regression of y_{i2} on y_{i1} .

More generally, (7.18) is also ML for $\hat{\mu}_2$ if the data are MAR and

1. y_{i2} given y_{i1} is normal with mean $(\beta_{20.1} + y_{i1}\beta_{21.1})$ and variance $\sigma_{22.1}$.
2. y_{i1} has any distribution such that (i) $\hat{\mu}_1$ is the ML estimate of the mean of y_{i1} , and (ii) μ_1 and $\beta_{20.1}$, $\beta_{21.1}$, $\sigma_{22.1}$ are distinct from the parameters of the distribution of y_{i1} .

A special case is *dummy variable regression*, where y_{i1} represents K dummy variables indicating $K+1$ groups. The k th component of y_{i1} is defined to be 1 if the i th unit belongs to group k and is zero otherwise: For units in group 1, $y_{i1} = (1, 0, 0, \dots, 0)$; for units in group 2, $y_{i1} = (0, 1, 0, \dots, 0)$; for units in group K , $y_{i1} = (0, 0, \dots, 0, 1)$; and for units in group $K+1$, $y_{i1} = (0, 0, \dots, 0, 0)$; group $K+1$ is often called the *reference group*.

With these definitions, $\hat{\mu}_1$ is a vector consisting of the proportions of the n sampled units in each of the first K groups, μ_1 is the corresponding vector of expected proportions, and aforementioned condition 2 is satisfied. Condition 1 is equivalent to assuming that all values of y_{i2} in group k are normal with the same group mean and same variance $\sigma_{22.1}$.

By the properties of dummy variable regression, the predicted value of y_{i2} for a unit in group k is the mean of the observed values of y_{i2} in group k . Thus, the ML estimate corresponds to imputing the group means for missing values of y_{i2} , a form of mean imputation that we discussed when considering nonresponse in sample surveys in Chapter 4.

7.4.2 The Factored Likelihood for a General Monotone Pattern

The methods described in Sections 7.2 and 7.3 can be readily generalized to the monotone pattern of data in Figure 1.1c, where for unit i , y_{ij} is recorded if $y_{i,j+1}$ is recorded ($j = 1, \dots, J-1$), so that Y_1 is more observed than Y_2 , which

is more observed than Y_3 , and so on (Rubin 1974). We confine attention to ML estimation. Precision of estimation and Bayesian inference can be addressed using straightforward extensions of the methods of Sections 7.2.2 and 7.3.

The appropriate factored likelihood for this pattern is

$$\begin{aligned} & \prod_{i=1}^n f_Y(y_{i1}, \dots, y_{ij} | \phi) \\ &= \prod_{i=1}^n f_1(y_{i1} | \phi_1) \prod_{i=1}^{r_2} f_2(y_{i2} | y_{i1}, \phi_2) \cdots \prod_{i=1}^{r_j} f_j(y_{ij} | y_{i1}, \dots, y_{i,j-1}, \phi_j), \end{aligned}$$

where for $j = 1, \dots, J$, $f_j(y_{ij} | y_{i1}, \dots, y_{i,j-1}, \phi_j)$ is the conditional distribution of y_{ij} given $y_{i1}, \dots, y_{i,j-1}$, indexed by the parameter ϕ_j . If (y_{i1}, \dots, y_{ij}) follows a multivariate normal distribution, then $f_j(y_{ij} | y_{i1}, \dots, y_{i,j-1}, \phi_j)$ is a normal distribution with mean linear in $y_{i1}, \dots, y_{i,j-1}$ and with constant variance. With the usual unrestricted natural parameter space of ϕ , the ϕ_j are distinct, and so ML estimates of ϕ_j are obtained by regressing the y_{ij} on the $y_{i1}, \dots, y_{i,j-1}$, using the set of units for which y_{i1}, \dots, y_{ij} are all observed.

Example 7.5 Multivariate Normal Monotone Data. Marini et al. (1980) provide a numerical illustration of ML estimation for monotone data with $J > 2$ patterns, for a panel study with 4352 units. The data pattern, given in Table 1.1, is not monotone, but as noted in Chapter 1, a monotone pattern can be achieved by discarding some data, in particular, those superscripted by the letter b in the table. The resulting pattern is monotone as in Figure 1.1c with $J = 4$. Assuming normality, ML estimates of the mean and covariance matrix of the variables can be found by the following procedure:

1. Calculate the mean vector and covariance matrix for the fully observed block 1 variables, from all the units.
2. Calculate the multivariate linear regression of the next most observed variables, block 2 variables on block 1 variables, from units with both block 1 and block 2 variables recorded.
3. Calculate the multivariate linear regression of block 3 variables on block 1 and 2 variables, from units with block 1–3 variables recorded.
4. Calculate the multivariate linear regression of block 4 variables on block 1–3 variables from units with all variables recorded.

ML estimates of the means and covariance matrix of all the variables can be obtained as functions of the parameter estimates in 1–4. The computational details, involving the powerful sweep operator, are discussed in the next section. Results are shown in Table 7.3.

The first column of Table 7.3 gives a description of the variables. The next two columns give ML estimates $\hat{\mu}_{\text{ML}}$ of the means and $\hat{\sigma}_{\text{ML}}$ of the standard

Table 7.3 Example 7.5: maximum likelihood estimates of means and standard deviations obtained for the entire original sample and comparisons with two alternative sets of estimates

Variable	Units		Maximum likelihood estimates				Estimates based on available units				Estimates based on complete units			
	Mean	SD	Mean	SD	$\frac{(\hat{\mu}_A - \hat{\mu}_{ML})}{\hat{\sigma}_{ML}} \times 100$	$\frac{(\hat{\sigma}_A - \hat{\sigma}_{ML})}{\hat{\sigma}_{ML}} \times 100$	Mean	SD	$\frac{(\hat{\mu}_{CC} - \hat{\mu}_{ML})}{\hat{\sigma}_{ML}} \times 100$	$\frac{(\hat{\sigma}_{CC} - \hat{\sigma}_{ML})}{\hat{\sigma}_{ML}} \times 100$	$\frac{(\hat{\sigma}_{CC} - \hat{\sigma}_{ML})}{\hat{\sigma}_{ML}} \times 100$			
<i>Block I variables: measured during adolescence</i>														
Father's education	11.7	3.5	11.7	3.5	0	0	12.10	3.4	9.9	-1.6				
Mother's education	11.5	2.9	11.51	2.9	0	0	11.9	2.9	12.1	-2.4				
Father's occupation	6.12	2.90	6.12	2.90	0	0	6.41	2.87	10.1	-1.2				
Intelligence	106.6	12.9	106.6	12.9	0	0	109.0	11.2	18.7	-13.4				
College preparatory curriculum	0.41	0.49	0.41	0.49	0	0	0.53	0.50	13.4	1.4				
Time spent on homework	1.59	0.81	1.59	0.81	0	0	1.63	0.80	5.4	-2.3				
Grade point average	2.32	0.77	2.32	0.77	0	0	2.59	0.70	34.9	-9.3				
College plans	0.49	0.50	0.49	0.50	0	0	0.60	0.49	21.4	-1.8				
Friends' college plans	0.51	0.37	0.51	0.37	0	0	0.57	0.35	16.3	-4.1				
Participation in extracurricular activities	0.41	0.49	0.41	0.49	0	0	0.49	0.50	15.8	1.4				
Membership in top leading crowd	0.09	0.28	0.09	0.28	0	0	0.13	0.34	8.6	19.4				
Membership in intermediate leading crowd	0.17	0.38	0.17	0.38	0	0	0.20	0.40	5.6	4.1				
Cooking/drinking	0.57	1.03	0.57	1.03	0	0	0.48	0.84	-8.4	-9.4				

Dating frequency at time of survey	4.03	4.80	4.03	4.80	0	0	3.70	4.52	-6.8	-5.8
Liking for self	2.37	0.53	2.37	0.53	0	0	2.36	0.52	-0.4	-1.9
Grade in school	2.43	1.05	2.43	1.05	0	0	2.50	1.06	6.1	1.5
<i>Block 2 variables: measured for all follow-up respondents</i>										
Educational attainment	13.6	2.3	13.3	2.3	5.5	-1.4	14.2	24.9	-4.0	
Occupational prestige	44.4	13.0	45.1	12.9	5.2	-0.9	47.1	12.7	20.4	-2.0
Marital status	0.94	0.24	0.94	0.24	0.0	0.0	0.94	0.24	0.0	-0.4
Number of children	1.99	1.31	1.97	1.30	-1.4	-0.2	1.93	1.24	-4.8	-4.9
Age	30.6	1.2	30.7	1.2	2.1	0.3	30.7	1.2	7.9	-5.4
Father's occupational prestige	44.0	14.8	44.3	14.8	1.8	-0.2	44.8	14.3	5.3	-3.2
<i>Block 3 variables: measured only for initial questionnaire</i>										
Respondents to the follow-up										
Personal esteem	3.13	0.38	3.15	0.38	5.2	0.3	3.15	0.37	5.3	-1.1
Dating frequency during last two years of high school	4.37	3.41	4.20	3.26	-5.1	-1.4	4.21	3.35	-4.7	-1.6
Number of siblings	2.22	1.75	2.10	1.74	-6.9	-0.2	2.06	1.66	-9.4	-5.0
<i>Block 4 variables: measured on parents' questionnaire</i>										
Family income	4.09	1.53	4.08	1.54	-1.1	0.5	4.22	1.57	8.0	2.6
Parental encouragement to go to college	0.71	0.43	0.71	0.46	-1.6	4.8	0.75	0.43	8.0	-0.7
Number of children in family of origin	3.04	1.54	3.07	1.67	1.8	8.6	2.98	1.55	-4.2	0.8

deviations of each variable. The next two columns display estimates ($\hat{\mu}_A, \hat{\sigma}_A$) from the available cases method (that is, the sample means and standard deviations using all the units available for each variable, as in Section 3.4). The two columns after these estimates display the magnitude of differences between the ML and available-case estimates, measured in percent standard deviations estimated by ML. Even though these estimates of marginal parameters are quite close to the ML estimates, the available-case method is not generally recommended, as discussed in Chapter 3. Finally, the last four columns of the table present and compare with ML the estimates ($\hat{\mu}_{CC}, \hat{\sigma}_{CC}$) based only on the 1594 complete units, the complete-case method discussed in Chapter 3. Estimates of means from this procedure sometimes differ markedly from the ML estimates. For example, the estimate for grade-point average is 0.35 of a standard deviation higher than the ML estimate, indicating that students lost to follow-up appear to have lower scores than average.

7.4.3 ML Computation for Monotone Normal Data via the Sweep Operator

We now review the use of the *sweep operator* (Beaton 1964) in linear regression with complete units and show how this operator provides a simple and convenient way to perform the ML calculations for incomplete normal data. The version of sweep we describe is not exactly the one originally defined in Beaton (1964); rather, it is the one defined by Dempster (1969); another accessible reference is Goodnight (1979). The sweep operator will also be useful in Chapter 9 when we consider ML estimation for normal data with a general pattern of missingness.

The sweep operator is defined for symmetric matrices as follows. The $p \times p$ symmetric matrix G is said to be *swept on row and column k* if it is replaced by another symmetric $p \times p$ matrix H with elements defined as follows:

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk}, \quad j \neq k, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk}, \quad j \neq k, \quad l \neq k. \end{aligned} \tag{7.19}$$

To illustrate (7.19), consider the 3×3 case:

$$G = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{12} & g_{22} & g_{23} \\ g_{13} & g_{23} & g_{33} \end{bmatrix},$$

$$H = \text{SWP}[1]G = \begin{bmatrix} -1/g_{11} & g_{12}/g_{11} & g_{13}/g_{11} \\ g_{12}/g_{11} & g_{22} - g_{12}^2/g_{11} & g_{23} - g_{13}g_{12}/g_{11} \\ g_{13}/g_{11} & g_{23} - g_{13}g_{12}/g_{11} & g_{33} - g_{13}^2/g_{11} \end{bmatrix}.$$

We use the notation $\text{SWP}[k]G$ to denote the matrix H defined by (7.19). Also, the result of successively applying the operations $\text{SWP}[k_1]$, $\text{SWP}[k_2], \dots$, $\text{SWP}[k_t]$ to the matrix G will be denoted by $\text{SWP}[k_1, k_2, \dots, k_t]G$. In actual computations, the sweep operation is efficiently achieved by first replacing g_{kk} by $h_{kk} = -1/g_{kk}$, then replacing the remaining elements g_{jk} and g_{kl} in row and column k by $h_{jk} = h_{kj} = -g_{jk}h_{kk}$, and finally replacing elements g_{il} that are neither in row k nor in column k by $h_{jl} = g_{jl} - h_{jk}g_{kl}$. Storage space can be saved by storing the distinct elements of the symmetric $p \times p$ matrices in vectors of length $p(p+1)/2$, so that for $k \leq j$, the (j, k) th element of the matrix is stored as the $(j(j-1)/2 + k)$ th element of the vector.

Some algebra shows that the sweep operator is commutative, that is

$$\text{SWP}[j, k]G = \text{SWP}[k, j]G.$$

It follows more generally that

$$\text{SWP}[j_1, \dots, j_t]G = \text{SWP}[k_1, \dots, k_t]G,$$

where j_1, \dots, j_t is any permutation of k_1, \dots, k_t . That is, the order in which a set of sweeps is carried out does not affect the final answer algebraically, although some orders may be computationally more accurate than others.

The sweep operator is closely related to linear regression. For example, suppose that G is a (2×2) covariance matrix of two variables Y_1 and Y_2 , and let $H = \text{SWP}[1]G$. Then h_{12} is the coefficient of Y_1 from the regression of Y_2 on Y_1 , and h_{22} is the residual variance of Y_2 given Y_1 . Furthermore, when G is a sample covariance matrix from n independent units, $-h_{11}h_{22}/n$ is the estimated sampling variance of the sample regression coefficient, h_{12} .

More generally, suppose we have a sample of n units on K variables Y_1, \dots, Y_K . Let G denote the $(K+1) \times (K+1)$ matrix

$$G = \begin{bmatrix} 1 & \bar{y}_1 & \cdots & \bar{y}_j & \cdots & \bar{y}_K \\ \bar{y}_1 & n^{-1} \sum y_1^2 & & & & n^{-1} \sum y_K y_1 \\ \vdots & \vdots & \ddots & & & \vdots \\ \bar{y}_k & & & n^{-1} \sum y_j y_k & & \\ \bar{y}_K & n^{-1} \sum y_1 y_K & & & \ddots & \\ & & & & & n^{-1} \sum y_K^2 \end{bmatrix},$$

where $\bar{y}_1, \dots, \bar{y}_K$ are the sample means, and summations are over the n units. For convenience, we index the rows and columns from 0 to K , so that row and column j corresponds to variable Y_j . Sweeping on row and column 0 yields

$$\text{SWP}[0]G = \begin{bmatrix} -1 & \bar{y}_1 & \cdots & \bar{y}_j & \cdots & \bar{y}_K \\ \bar{y}_1 & \hat{\sigma}_{11} & & & \cdots & \hat{\sigma}_{K1} \\ \vdots & \vdots & \ddots & & & \vdots \\ \bar{y}_k & & & \hat{\sigma}_{jk} & & \\ & & & & \ddots & \\ \bar{y}_K & \hat{\sigma}_{1K} & & \cdots & & \hat{\sigma}_{KK} \end{bmatrix}, \quad (7.20)$$

where $\hat{\sigma}_{jk}$ is the sample covariance of Y_j and Y_k , with denominator n rather than $(n - 1)$. This operation corresponds to correcting the scaled cross-products matrix of Y_1, \dots, Y_K , G , for the means of Y_1, \dots, Y_K to create the covariance matrix. In terms of regression, the means in the first row and column of $\text{SWP}[0]G$ are coefficients from the regression of Y_1, \dots, Y_K on the constant term $Y_0 \equiv 1$, and the corrected, scaled cross-products matrix $\{\hat{\sigma}_{jk}\}$ is the residual covariance matrix (with divisor n rather than $n - 1$) from this regression. Thus, we also call this process sweeping on the constant term. We call (7.20) the *augmented covariance matrix* of the variables Y_1, \dots, Y_K .

Sweeping (7.20) on row and column 1, corresponding to Y_1 , yields the symmetric matrix

$$\begin{aligned} \text{SWP}[0, 1]G &= \begin{bmatrix} -(1 + \bar{y}_1/\hat{\sigma}_{11}) & \bar{y}_{11}/\hat{\sigma}_{11} & \bar{y}_2 - (\hat{\sigma}_{12}/\hat{\sigma}_{11})\bar{y}_1 & \cdots & \bar{y}_K - (\hat{\sigma}_{1K}/\hat{\sigma}_{11})\bar{y}_1 \\ -1/\hat{\sigma}_{11} & \hat{\sigma}_{12}/\hat{\sigma}_{11} & \cdots & & \hat{\sigma}_{1K}/\hat{\sigma}_{11} \\ & \hat{\sigma}_{22} - \hat{\sigma}_{12}^2/\hat{\sigma}_{11} & \cdots & \hat{\sigma}_{2K} - \hat{\sigma}_{1K}\hat{\sigma}_{12}/\hat{\sigma}_{11} & \\ & & & & \vdots \\ \bar{y}_K - (\hat{\sigma}_{1K}/\hat{\sigma}_{11})\bar{y}_1 & & & & \hat{\sigma}_{KK} - \hat{\sigma}_{1K}^2/\hat{\sigma}_{11} \end{bmatrix} \\ &= \begin{bmatrix} -A & B \\ B^T & C \end{bmatrix}, \end{aligned}$$

say, where A is (2×2) , B is $2 \times (K - 1)$, and C is $(K - 1) \times (K - 1)$. This matrix yields results for the (multivariate) regression of Y_2, \dots, Y_K on Y_1 . In particular, the j th column of B gives the intercept and slope for the regression of Y_{j+1} on Y_1 for $j = 1, \dots, K - 1$. The matrix C gives the residual covariance matrix of Y_2, \dots, Y_K given Y_1 . Finally, the elements of A , when multiplied by the appropriate residual variance or covariance in C and divided by n , yield the asymptotic sampling variances and covariances of the estimated regression coefficients in B .

Sweeping the constant term and the first q elements yields results for the multivariate regression of Y_{q+1}, \dots, Y_K on Y_1, \dots, Y_q . Specifically, letting

$$\text{SWP}[0, 1, \dots, q]G = \begin{bmatrix} -D & E \\ E^T & F \end{bmatrix},$$

where D is $(q+1) \times (q+1)$, E is $(q+1) \times (K-q)$, and F is $(K-q) \times (K-q)$, the j th column of E gives the least squares intercept and slopes of the regression of Y_{j+q} on Y_1, \dots, Y_q , for $j = 1, 2, \dots, K-q$; the matrix F is the residual covariance matrix of Y_{q+1}, \dots, Y_K ; and the elements of D can be used, as previously mentioned, to give the asymptotic sampling variances and covariances of the estimated regression coefficients in E .

In summary, ML estimates for the multivariate linear regression of Y_{q+1}, \dots, Y_K on Y_1, \dots, Y_q can be found by sweeping the rows and columns corresponding to the constant term and the predictor variables Y_1, \dots, Y_q from the scaled cross-products matrix G .

The operation of sweeping a variable in effect turns that variable from an outcome (or dependent) variable into a predictor (or independent) variable. There is also an operator inverse to sweep that turns predictor variables into outcome variables. This operator is called *reverse sweep* (RSW) and is defined by

$$H = \text{RSW}[k]G,$$

where

$$\begin{aligned} h_{kk} &= -1/g_{kk}, \\ h_{jk} &= h_{kj} = -g_{jk}/g_{kk}, \quad j \neq k, \\ h_{jl} &= g_{jl} - g_{jk}g_{kl}/g_{kk}, \quad j \neq k, \quad l \neq k. \end{aligned} \tag{7.21}$$

It is readily verified that reverse sweep is also commutative and is the inverse operator to sweep; that is

$$(\text{RSW}[k])(\text{SWP}[k])G = (\text{SWP}[k])(\text{RSW}[k])G = G.$$

Example 7.6 Bivariate Normal Monotone Data (Example 7.1 Continued). Various parameterizations of the bivariate normal distribution are easily related using the sweep and reverse sweep operators. Thus, the parameters θ and ϕ of Example 7.1 and the relationships in (7.4) and (7.5) can be compactly expressed using the SWP[] and RSW[] notation. Also, numerical values of standard functions of ML estimates can be simply computed using these operators.

Suppose we arrange $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$ in the following (3×3) symmetric matrix, which is the population analog of (7.20):

$$\theta^* = \begin{bmatrix} -1 & \mu_1 & \mu_2 \\ \mu_1 & \sigma_{11} & \sigma_{12} \\ \mu_2 & \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

The matrix θ^* represents the parameters of the bivariate normal with the constant swept. If θ^* is swept on row and column 1, then we obtain from (7.19)

$$\text{SWP}[1]\theta^* = \begin{bmatrix} -(1 + \mu_1^2/\sigma_{11}) & \mu_1/\sigma_{11} & \mu_2 - \mu_1\sigma_{12}/\sigma_{11} \\ \mu_1/\sigma_{11} & -\sigma_{11}^{-1} & \sigma_{12}/\sigma_{11} \\ \mu_2 - \mu_1\sigma_{12}/\sigma_{11} & \sigma_{12}/\sigma_{11} & \sigma_{22} - \sigma_{12}^2/\sigma_{11} \end{bmatrix}.$$

An examination of (7.4) reveals that row (or column) 2 of $\text{SWP}[1]\theta^*$ provides the intercept, $\mu_2 - \mu_1\sigma_{12}/\sigma_{11}$; the slope of the regression of Y_2 on Y_1 , σ_{12}/σ_{11} ; and the residual variance, $\sigma_{22-1} = \sigma_{22} - \sigma_{12}^2/\sigma_{11}$. Also, although not in a particularly familiar form, the 2×2 submatrix formed by the first two rows and columns provides the parameters of the marginal distribution of Y_1 . To see this, write

$$\phi^* = \text{SWP}[1]\theta^* = \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \mu_1 \\ \mu_1 & \sigma_{11} \end{bmatrix} & \beta_{20-1} \\ & \begin{bmatrix} \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \hat{\beta}_{21-1} \\ & \beta_{20-1} & \hat{\sigma}_{22-1} \end{bmatrix}, \quad (7.22)$$

where ϕ^* is a slightly modified version of $\phi = (\mu_1, \sigma_{11}, \beta_{20-1}, \beta_{21-1}, \sigma_{22-1})^T$, displayed as a matrix. By Property 6.1, a similar expression relates the ML estimates of θ to the ML estimates of ϕ :

$$\hat{\phi}^* = \text{SWP}[1]\hat{\theta}^* = \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \hat{\beta}_{20-1} \\ & \begin{bmatrix} \hat{\beta}_{21-1} \end{bmatrix} & \hat{\sigma}_{22-1} \end{bmatrix}.$$

Applying the $\text{RSW}[1]$ operator to both sides yields

$$\hat{\theta}^* = \text{RSW}[1] \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1 \\ \hat{\mu}_1 & \hat{\sigma}_{11} \end{bmatrix} & \hat{\beta}_{20-1} \\ & \begin{bmatrix} \hat{\beta}_{21-1} \end{bmatrix} & \hat{\sigma}_{22-1} \end{bmatrix}. \quad (7.23)$$

Expression (7.23) defines the transformation from $\hat{\phi}$ to $\hat{\theta}$ in terms of the sweep and reverse sweep operators and thus shows how these operators can be used to compute $\hat{\theta}$ from $\hat{\phi}$.

Example 7.7 Multivariate Normal Monotone Data (Example 7.5 Continued). We now extend Example 7.6 to show how the sweep and reverse sweep operators can be applied to find ML estimates of the mean and covariance matrix of a multivariate normal distribution from data with a monotone pattern. We assume that the data have the monotone pattern of Figure 1.1c, after suitable

arrangement of the variables. Also for simplicity, we consider the case with $J = 3$ blocks of variables. The extension to more than three blocks of variables is immediate.

Step 1: Find the ML estimates $\hat{\mu}_1$ and $\hat{\Sigma}_{11}$ of the mean μ_1 and covariance matrix Σ_{11} of the first block of variables, which are completely observed. These are simply the sample mean and sample covariance matrix of Y_1 based on all the units.

Step 2: Find the ML estimates $\hat{\beta}_{20 \cdot 1}$, $\hat{\beta}_{21 \cdot 1}$, and $\hat{\Sigma}_{22 \cdot 1}$ of the intercepts, regression coefficients, and residual covariance matrix for the regression of Y_2 on Y_1 . These can be found by sweeping the variables Y_1 out of the augmented covariance matrix of Y_1 and Y_2 based on the units with Y_1 and Y_2 both observed.

Step 3: Find the ML estimates $\hat{\beta}_{30 \cdot 12}$, $\hat{\beta}_{31 \cdot 12}$, $\hat{\beta}_{32 \cdot 12}$, and $\hat{\Sigma}_{33 \cdot 12}$ of the intercepts, regression coefficients, and residual covariance matrix for the regression of Y_3 on Y_1 and Y_2 . These can be found by sweeping the variables Y_1 and Y_2 out of the augmented covariance matrix of Y_1 , Y_2 , and Y_3 based on the units with Y_1 , Y_2 , and Y_3 observed.

Step 4: Calculate the matrix

$$A = \text{SWP}[1] \begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & A_{22} \end{bmatrix},$$

where $\text{SWP}[1]$ is shorthand for sweeping on the set of variables Y_1 .

Step 5: Calculate the matrix

$$B = \text{SWP}[2] \begin{bmatrix} a_{11} & a_{12} & \hat{\beta}_{20 \cdot 1}^T \\ a_{21} & A_{22} & \hat{\beta}_{21 \cdot 1}^T \\ \hat{\beta}_{20 \cdot 1} & \hat{\beta}_{21 \cdot 1} & \hat{\Sigma}_{22 \cdot 1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix},$$

where $\text{SWP}[2]$ is shorthand for sweeping on the set of variables Y_2 .

Step 6: Finally, the ML estimate of the augmented covariance matrix of Y_1 , Y_2 , and Y_3 is given by

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \text{RSW}[1, 2] \begin{bmatrix} c_{11} & c_{12} & c_{13} & \hat{\beta}_{20 \cdot 1}^T \\ c_{21} & c_{22} & c_{23} & \hat{\beta}_{31 \cdot 12}^T \\ c_{31} & c_{32} & c_{33} & \hat{\beta}_{32 \cdot 12}^T \\ \hat{\beta}_{20 \cdot 1} & \hat{\beta}_{31 \cdot 12} & \hat{\beta}_{32 \cdot 12} & \hat{\Sigma}_{33 \cdot 12} \end{bmatrix}.$$

This matrix contains the ML estimates of the mean and covariance matrix of Y_1 , Y_2 , and Y_3 , as indicated.

Steps 4–6 can be represented concisely by the equation

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \text{RSW}[1, 2] \begin{bmatrix} \text{SWP}[1] & \begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} & \begin{bmatrix} \hat{\beta}_{20 \cdot 1}^T \\ \hat{\beta}_{21 \cdot 1}^T \end{bmatrix} & \hat{\beta}_{30 \cdot 12}^T \\ \text{SWP}[2] & \begin{bmatrix} \hat{\beta}_{20 \cdot 1}^T & \hat{\beta}_{21 \cdot 1}^T & \hat{\beta}_{31 \cdot 12}^T \\ \hat{\beta}_{21 \cdot 1}^T & \hat{\Sigma}_{22 \cdot 1} & \hat{\beta}_{32 \cdot 12}^T \\ \hat{\beta}_{30 \cdot 12}^T & \hat{\beta}_{31 \cdot 12}^T & \hat{\Sigma}_{33 \cdot 12} \end{bmatrix} \end{bmatrix},$$

with obvious generalizations to more than three blocks of variables. This equation defines the transformation from $\hat{\phi}$ to $\hat{\theta}$ for this situation.

Estimates of the sampling precision of the ML estimates based on the asymptotic covariance matrix are not as easily accomplished by such operations. However, a simple alternative approach is to adopt a Bayesian perspective and estimate the precision using the posterior variance, as introduced in Section 7.3 for bivariate normal monotone data. We discuss this approach in Section 7.4.4.

Example 7.8 A Numerical Example. Rubin (1976c) presents the previously described calculations for the data in Table 7.4, described in Draper and

Table 7.4 The data for Example 7.8^a

Unit	Variables				
	X_1	X_2	X_3	X_4	$Y = X_5$
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	(6)	102.7
8	1	31	22	(44)	72.5
9	2	54	18	(22)	93.1
10	(21)	(47)	4	(26)	115.9
11	(1)	(40)	23	(34)	83.8
12	(11)	(66)	9	(12)	113.3
13	(10)	(68)	8	(12)	109.4

^aValues in parentheses are considered missing in the example.

Smith (1981). The original labeling of the variables is X_1, \dots, X_5 . The data have the pattern of Figure 1.1c with $J = 3$ and $Y_1 = (X_3, X_5)$, $Y_2 = (X_1, X_2)$, and $Y_3 = X_4$. We first apply the method of Example 7.7 to yield ML estimates of the parameters.

Step 1 gives the ML estimates of the marginal distribution of (X_3, X_5) as

$$\hat{\mu}_3 = 11.769, \quad \hat{\mu}_5 = 95.423, \quad \hat{\sigma}_{33} = 37.870, \\ \hat{\sigma}_{35} = -47.566, \quad \hat{\sigma}_{55} = 208.905.$$

Step 2 gives ML estimates of the coefficients from the regression of (X_1, X_2) on (X_3, X_5) , based on units 1–9:

$$\hat{\beta}_{10\cdot 35} = 2.802, \quad \hat{\beta}_{13\cdot 35} = -0.526, \quad \hat{\beta}_{15\cdot 35} = 0.105, \\ \hat{\beta}_{20\cdot 35} = -74.938, \quad \hat{\beta}_{23\cdot 35} = 1.062, \quad \hat{\beta}_{25\cdot 35} = 1.178,$$

and the residual covariance matrix

$$\hat{\Sigma}_{12\cdot 35} = \begin{matrix} X_1 & X_2 \\ \begin{matrix} X_1 \\ X_2 \end{matrix} & \begin{bmatrix} 3.804 & -8.011 \\ -8.011 & 24.382 \end{bmatrix} \end{matrix}.$$

Step 3 gives ML estimates of the coefficients and residual variance for the regression of X_4 on the other variables, based on units 1–6:

$$\hat{\beta}_{40\cdot 1235} = 85.753, \quad \hat{\beta}_{41\cdot 1235} = -1.863, \quad \hat{\beta}_{42\cdot 1235} = -1.324, \\ \hat{\beta}_{43\cdot 1235} = -1.533, \quad \hat{\beta}_{45\cdot 1235} = 0.397, \quad \hat{\sigma}_{44\cdot 1235} = 0.046.$$

Steps 4–6 yield

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \text{RSW}[1235] \begin{bmatrix} \text{SWP}[35] & \begin{bmatrix} -1 & 11.769 & 95.423 \\ 11.769 & 37.870 & -47.566 \\ 95.423 & -47.566 & 208.905 \end{bmatrix} & \begin{bmatrix} 2.802 & -74.938 \\ -0.526 & 1.062 \\ 0.105 & 1.178 \end{bmatrix} & \begin{bmatrix} 85.753 \\ -1.533 \\ 0.397 \end{bmatrix} \\ \text{SWP}[12] & \begin{bmatrix} 2.802 & -0.526 & 0.105 & 3.804 & -8.011 & -1.863 \\ -74.938 & 1.062 & 1.178 & -8.011 & 24.382 & -1.324 \\ 85.753 & -1.533 & 0.397 & -1.863 & -1.324 & 0.046 \end{bmatrix} \end{bmatrix}$$

Calculating the right side and reordering the variables gives ML estimates

$$\hat{\mu}^T = [6.655 \quad 49.965 \quad 11.769 \quad 27.047 \quad 95.423]$$

$$\hat{\Sigma} = \begin{bmatrix} 21.826 & 20.864 & -24.900 & -11.473 & 46.953 \\ 20.864 & 238.012 & -15.817 & -252.072 & 195.604 \\ -24.900 & -15.817 & 37.870 & -9.599 & -47.556 \\ -11.473 & -252.072 & -9.599 & 294.183 & -190.599 \\ 46.953 & 195.604 & -47.556 & -190.599 & 208.905 \end{bmatrix}.$$

7.4.4 Bayes Computation for Monotone Normal Data via the Sweep Operator

In the situation of Example 7.7, Bayesian inference is achieved by replacing the ML estimates of ϕ in steps 1–3, namely

$$\hat{\phi} = (\hat{\mu}_1, \hat{\Sigma}_{11}, \hat{\beta}_{20\cdot 1}, \hat{\beta}_{21\cdot 1}, \hat{\Sigma}_{22\cdot 1}, \hat{\beta}_{30\cdot 12}, \hat{\beta}_{31\cdot 12}, \hat{\beta}_{32\cdot 12}, \hat{\Sigma}_{33\cdot 12}),$$

by D draws from the posterior distribution of ϕ :

$$\phi^{(d)} = (\mu_1^{(d)}, \Sigma_{11}^{(d)}, \beta_{20\cdot 1}^{(d)}, \beta_{21\cdot 1}^{(d)}, \Sigma_{22\cdot 1}^{(d)}, \beta_{30\cdot 12}^{(d)}, \beta_{31\cdot 12}^{(d)}, \beta_{32\cdot 12}^{(d)}, \Sigma_{33\cdot 12}^{(d)}), \quad d = 1, \dots, D,$$

and then applying the sweep operations in steps 4–6 to obtain D draws $\theta^{(d)}$ from the posterior distribution of θ , $d = 1, \dots, D$. These draws can then be used to simulate the posterior distribution of θ and thereby produce interval estimates for all the parameters.

Example 7.9 *Inferences for Data in Example 7.8.* For comparison purposes, the bootstrap was applied to the data in Example 7.8, yielding the following means and standard errors over 1000 bootstrap samples:

$$\begin{aligned} x_1 &\quad x_2 &\quad x_3 &\quad x_4 &\quad x_5 \\ \text{Bootstrap means} &= [7.22 \quad 46.75 \quad 10.78 \quad 31.28 \quad 94.15] \\ \text{Bootstrap SEs} &= [1.10 \quad 3.14 \quad 1.35 \quad 3.42 \quad 2.97]. \end{aligned}$$

In contrast, the following are posterior means and standard deviations obtained as draws from the posterior distribution, using the Jeffreys' prior (6.35):

$$\begin{aligned} x_1 &\quad x_2 &\quad x_3 &\quad x_4 &\quad x_5 \\ \text{Posterior means} &= [6.73 \quad 49.93 \quad 11.66 \quad 27.14 \quad 95.51] \\ \text{Posterior SDs} &= [2.13 \quad 6.86 \quad 2.49 \quad 7.28 \quad 5.98]. \end{aligned}$$

We might expect the bootstrap means and the posterior means to be close to the ML estimates. This is true for the posterior means, but the bootstrap means are quite different from the ML estimates (e.g., 7.22 vs. 6.66 for the mean of x_1). More striking, the bootstrap standard errors are much smaller than the posterior standard deviations. The latter can be expected to be a bit larger in that they incorporate t -type corrections for estimating Σ that are not reflected in the bootstrap standard errors; however, this difference does not account for the large disparity. A more likely explanation is that the bootstrap standard errors are incorrect in this case, in view of the small sample size: note that there are only six complete units and five variables, the minimum required to attain unique ML estimates (Rubin 1994). Indeed, only about 5% of the bootstrap samples (1000 of 23 128) were included in the bootstrap calculation because the other samples did not yield unique parameter estimates. Our conclusion is that the simple bootstrap should not be used in such small-sample situations.

7.5 Factored Likelihoods for Special Nonmonotone Patterns

Nonmonotone patterns of incomplete data where factored likelihoods are possible have been noted by Anderson (1957), where each factor is a complete data likelihood, and the data are normal; Rubin (1974) generalizes this idea. The basic case is given by Figure 7.1. It has variables arranged into three blocks (Y_1 , Y_2 , Y_3) such that

1. Y_3 is *more observed* than Y_1 in the sense that for any unit for which Y_1 is at least partially observed, Y_3 is fully observed.
2. Y_1 and Y_2 are *never jointly observed*, in the sense that for any unit for which Y_2 is at least partially observed, Y_1 is entirely missing, and vice versa.
3. The rows of Y_1 are conditionally independent given Y_3 with the same parameters.

When Y_2 is ignored and Y_1 and Y_3 are scalar, Figure 7.1 reduces to bivariate monotone data. Under MAR, the loglikelihood of the data decomposes into

	Y_3	Y_2		Y_1		
Units ↓	0	...	0	1	...	1
	:	..	:	:	..	:
	0	...	0	1	...	1
	\times	...	\times	\times	...	\times
	:	..	:	:	..	:
	\times	...	\times	\times	...	\times

Figure 7.1 Data pattern where Y_3 is more observed than Y_1 , and Y_1 and Y_2 are never jointly observed; 0 = observed, 1 = missing, \times = possibly observed.
Source: Adapted from Rubin (1974).

two factors, one for the marginal distribution of Y_2 and Y_3 with parameter ϕ_{23} , based on all the units; and a second factor for the conditional distribution of Y_1 given Y_3 with parameter $\phi_{1\cdot 3}$, based on units with Y_3 fully observed. The proof of this result, which encompasses a proof of factorizations for monotone data, is given in Rubin (1974, Section 2).

The parameters ϕ_{23} and $\phi_{1\cdot 3}$ are often distinct because ϕ_{23} can be reparameterized (in the obvious notation) as $\phi_{2\cdot 3}$ and ϕ_3 , and the parameters $\phi_{1\cdot 3}$, $\phi_{2\cdot 3}$, and ϕ_3 are often distinct. An important aspect of this example is that ϕ_{23} and $\phi_{1\cdot 3}$ do not provide a complete reparametrization of the parameters of the joint distribution of Y_1 , Y_2 , and Y_3 because the parameters of conditional association (e.g., partial correlation) between Y_1 and Y_2 given Y_3 are not included in the aforementioned description. These parameters do not appear in the loglikelihood and cannot be estimated from the data.

Rubin (1974) shows how repeated reductions of the pattern of Figure 7.1 can be used to factorize the likelihood as fully as possible. Although in general not all the resultant factors can be dealt with independently using complete data methods, we illustrate the main ideas using two examples that do reduce to complete-data problems.

Example 7.10 A Normal Three-Variable Example. Lord (1955) and Anderson (1957) consider a trivariate normal sample with the pattern of Figure 7.1, with each of Y_1 , Y_2 , and Y_3 univariate, no completely observed units, r_1 units with Y_1 and Y_3 observed, and r_2 units with Y_2 and Y_3 observed, and $n = r_1 + r_2$. Assuming the data are MAR, the likelihood factors into three components: (i) $r_1 + r_2$ units with observations on the marginal normal distribution of Y_3 , with parameters μ_3 and σ_{33} ; (ii) r_1 units with observations on the conditional distribution of Y_1 given Y_3 , with intercept $\beta_{10\cdot 3}$, slope $\beta_{13\cdot 3}$, and variance $\sigma_{11\cdot 3}$; and (iii) r_2 units on the conditional distribution of Y_2 given Y_3 , with intercept $\beta_{20\cdot 3}$, slope $\beta_{23\cdot 3}$, and variance $\sigma_{22\cdot 3}$. These three components involve eight distinct parameters, whereas the original joint distribution of Y_1 , Y_2 , and Y_3 involves nine parameters, namely, three means, three variances, and three covariances. The omitted parameter in the reparametrization is the partial (conditional) correlation between Y_1 and Y_2 given Y_3 , about which there is no information in the observed data.

Data sets having such patterns of incompleteness are not uncommon. One context, where each Y_i is multivariate, is the *file-matching* problem, which arises when combining large government or medical databases. For example, suppose we have one file that is a random sample of Internal Revenue Service (IRS) records (with unit identifiers removed) and another file that is a random sample of Social Security Administration (SSA) records (also with unit identifiers removed, and no common units in the two files). The IRS file has detailed income information (Y_1) and background information (Y_3), whereas the SSA file has detailed work-history information (Y_2) and the same background

information (Y_3). The merged file can be viewed as a sample with Y_3 observed on all units but Y_1 and Y_2 never jointly observed. The term *file matching* is used to describe this situation because an attempt is often made to fill in the missing Y_1 and Y_2 values by matching units across files on the basis of Y_3 and imputing the missing values from matching units. Such problems are discussed in Rubin (1986) and Rässler (2002).

Example 7.11 An Application to Educational Data. In educational testing problems, such as given in Rubin and Thayer (1978), it is common that several new tests will be evaluated on different random samples from the same population. Specifically, let $X = (X_1, \dots, X_K)$ represent K standard tests given to all sampled subjects (that is, all units), and suppose new test Y_1 is given to the first sample of r_1 subjects, new test Y_2 is given to the second sample of r_2 subjects, and so on up to Y_q , where the samples have no subjects in common; because of the random sampling, the missing Y values are MCAR. Figure 7.2 displays the case with $q = 3$, which is a simple extension of the pattern in Example 7.10.

The partial correlations among the Y_j 's given X are inestimable in the strict sense that they do not have unique ML estimates. However, the simple correlations among the Y_j s are often of more interest in educational testing problems. Although these correlations do not have unique ML estimates, there is information in the data about their possible values.

Subsample	Unit	Standard tests			New tests		
		X_1, \dots, X_K			Y_1	Y_2	Y_3
1	1	0	...	0	0	1	1
	\vdots	\vdots		\vdots	\vdots	\vdots	\vdots
	r_1	0	...	0	0	1	1
2	$r_1 + 1$	0	...	0	1	0	1
	\vdots	\vdots		\vdots	\vdots	\vdots	\vdots
	$r_1 + r_2$	0	...	0	1	0	1
3	$r_1 + r_2 + 1$	0	...	0	1	1	0
	\vdots	\vdots		\vdots	\vdots	\vdots	\vdots
	$r_1 + r_2 + r_3$	0	...	0	1	1	0

Figure 7.2 Example 7.11: data structure with three new tests: 0 = score observed, 1 = score missing.

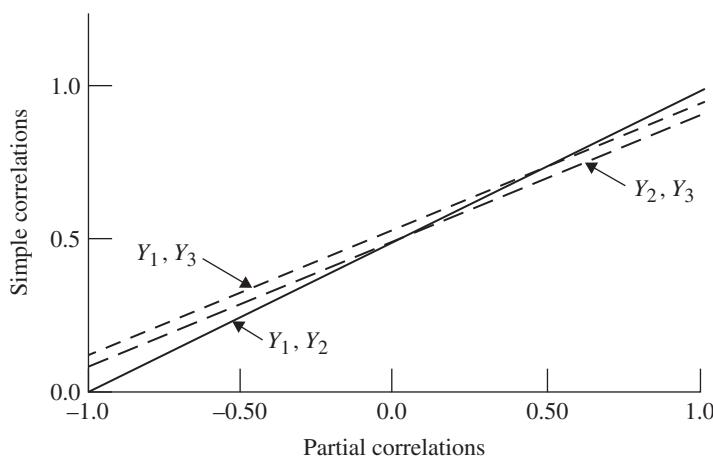


Figure 7.3 Example 7.11: simple correlations as functions of partial correlations. Source: Rubin and Thayer (1978). Reproduced with permission of Pschometrika.

Straightforward algebra shows that the simple correlation between Y_j and Y_k depends on the partial correlation between Y_j and Y_k but not on the partial correlation between any other pair of variables. As the partial correlation between Y_j and Y_k increases, the simple correlation between Y_j and Y_k increases; furthermore, this relationship is linear. Hence, given estimates of the simple correlation for two different values of the partial correlation (e.g., 0 and 1), one can estimate the simple correlation for any other value of the partial correlation using linear interpolation (or extrapolation, depending on the chosen values). Figure 7.3 displays plots of the estimated simple correlations as a function of the partial correlations for Education Testing Service data with the structure of Figure 7.2, with $r_1 = 1325$, $r_2 = 1345$, $r_3 = 2000$, and bivariate X (Rubin and Thayer 1978).

As with monotone normal data, the sweep operator is a useful notational and computational device for creating this figure. ML computations can be described as follows:

Step 1: Find the ML estimates of the marginal distribution of X , μ_x , and Σ_{xx} . These are simply the sample mean and covariance of all n units, $\hat{\mu}_x$ and $\hat{\Sigma}_{xx}$. This step yields $\hat{\mu}_x^T = (43.3, 26.8)$ and

$$\hat{\Sigma}_{xx} = \begin{bmatrix} 330.3 & 118.9 \\ 118.9 & 138.1 \end{bmatrix}.$$

Step 2: Find the ML estimates $\hat{\beta}_{10 \cdot x}$, $\hat{\beta}_{1x \cdot x}$, and $\hat{\sigma}_{11 \cdot x}^2$ of the regression coefficients and residual variance for the regression of Y_1 on X . These

can be found by sweeping the variables X out of the augmented sample covariance matrix of Y_1 and X based on the r_1 units with X and Y_1 both observed. This step yields $(\hat{\beta}_{10 \cdot x}, \hat{\beta}_{1x \cdot x}^T) = (0.99, 0.10, 0.17)$ and $\hat{\sigma}_{11 \cdot x} = 11.09$.

Step 3: Find the ML estimates $\hat{\beta}_{20 \cdot x}$, $\hat{\beta}_{2x \cdot x}$, and $\hat{\sigma}_{22 \cdot x}$ of the regression coefficients and residual variance for the regression of Y_2 on X . These can be found by sweeping the variables X out of the augmented sample covariance matrix of Y_2 and X based on the r_2 units with X and Y_2 both observed. This step yields $(\hat{\beta}_{20 \cdot x}, \hat{\beta}_{2x \cdot x}^T) = (-0.44, 0.18, 0.23)$ and $\hat{\sigma}_{22 \cdot x} = 27.38$.

Step 4: Find the ML estimates $\hat{\beta}_{30 \cdot x}$, $\hat{\beta}_{3x \cdot x}$, and $\hat{\sigma}_{33 \cdot x}$ of the regression coefficients and residual variance for the regression of Y_3 on X . These can be found by sweeping the variables X out of the augmented sample covariance matrix of Y_3 and X based on the r_3 units with X and Y_3 both observed. This step yields $(\hat{\beta}_{30 \cdot x}, \hat{\beta}_{3x \cdot x}^T) = (0.33, 0.23, 0.57)$ and $\hat{\sigma}_{33 \cdot x} = 71.49$.

Step 5: Fix all inestimable partial correlations at zero; then find unique ML estimates of the mean vector $\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

of all variables, as follows:

$$\begin{bmatrix} -1 & \hat{\mu}_x^T \\ \hat{\mu}_{(0)} & \hat{\Sigma}_{(0)} \end{bmatrix} = \text{RSW}[x] \begin{bmatrix} \text{SWP}[x] & \begin{bmatrix} -1 & \hat{\mu}_{xx}^T \\ \hat{\mu}_{xx} & \hat{\Sigma}_{xx} \end{bmatrix} & \begin{bmatrix} \hat{\beta}_{10 \cdot x} & \hat{\beta}_{20 \cdot x} & \hat{\beta}_{30 \cdot x} \\ \hat{\beta}_{1x \cdot x} & \hat{\beta}_{2x \cdot x} & \hat{\beta}_{3x \cdot x} \end{bmatrix} \\ \begin{bmatrix} \hat{\beta}_{10 \cdot x} & \hat{\beta}_{1x \cdot x}^T & \hat{\sigma}_{11 \cdot x} & 0 & 0 \\ \hat{\beta}_{20 \cdot x} & \hat{\beta}_{2x \cdot x}^T & 0 & \hat{\sigma}_{22 \cdot x} & 0 \\ \hat{\beta}_{30 \cdot x} & \hat{\beta}_{3x \cdot x}^T & 0 & 0 & \hat{\sigma}_{33 \cdot x} \end{bmatrix} \end{bmatrix}, \quad (7.24)$$

where the zeroes on the left side of (7.24) refer to the estimates being conditional on the zero partial correlations. Step 5 yields $\hat{\mu}_y^T = (9.96, 13.27, 25.63)$,

$$\hat{\Sigma}_{yy} = \begin{bmatrix} 22.66 & 17.61 & 32.84 \\ 17.61 & 54.31 & 49.61 \\ 32.84 & 49.61 & 165.64 \end{bmatrix},$$

$$\hat{\Sigma}_{xy} = \begin{bmatrix} 53.78 & 85.22 & 144.08 \\ 36.74 & 52.39 & 106.50 \end{bmatrix}.$$

Step 6: Fix all inestimable partial correlations at 1 and find the corresponding ML estimates

$$\begin{bmatrix} -1 & \hat{\mu}_{(1)}^T \\ \hat{\mu}_{(1)} & \hat{\Sigma}_{(1)} \end{bmatrix}.$$

These estimates are obtained by replacing the lower-right 3×3 submatrix on the right side of (7.24) with

$$\begin{bmatrix} \hat{\sigma}_{11.x} & \sqrt{\hat{\sigma}_{11.x}\hat{\sigma}_{22.x}} & \sqrt{\hat{\sigma}_{11.x}\hat{\sigma}_{33.x}} \\ \sqrt{\hat{\sigma}_{11.x}\hat{\sigma}_{22.x}} & \hat{\sigma}_{22.x} & \sqrt{\hat{\sigma}_{22.x}\hat{\sigma}_{33.x}} \\ \sqrt{\hat{\sigma}_{11.x}\hat{\sigma}_{33.x}} & \sqrt{\hat{\sigma}_{22.x}\hat{\sigma}_{33.x}} & \hat{\sigma}_{33.x} \end{bmatrix}.$$

This step yields the same value of $\hat{\mu}_y$ and the diagonal of $\hat{\Sigma}_{yy}$ but different estimates of the other parameters. In particular, the estimated correlations among the Y variables are 0.999, 0.996, and 0.990. The corresponding estimates from step 5 are 0.50, 0.54, and 0.52.

Linear interpolation between the correlations reported in steps 5 and 6 produces Figure 7.3. Other parameters, such as multiple correlations, were also considered in Rubin and Thayer (1978). In general, these are not linear in the partial correlations with no unique ML estimates, but they are still simple to compute.

Bayesian inferences for these two examples are obtained by replacing ML estimates of the parameters in the factored likelihood by draws, as in Section 7.4.4. The joint posterior distribution of the inestimable parameters is the same as their prior distribution, which has to be proper in these situations.

Example 7.12 *Correcting for Measurement Error Using External Calibration Data (Example 1.15 Continued).* As described in Example 1.15, Figure 1.4b displays data on four variables X , W , Y , Z from a main sample and an external calibration sample. Guo et al. (2011) discuss inference for the regression of vector outcome Y (dimension $q \geq 1$) on predictors X (scalar) and Z (dimension $r \geq 0$). Because q may be greater than one, the formulation covers multivariate regression with more than one dependent variable. In the main sample, X is completely missing, and Y and Z are completely observed, together with a variable W that is a proxy for X , subject to measurement error. In the calibration sample, pairs of X and W are obtained, typically collected independently of the main study, for example by an assay manufacturer if W represents a chemical assay.

It is convenient to combine Y and Z into a p -dimensional variable $U = (Y, Z)$, where $p = q + r$. We assume that

- (a) the missing data in Figure 1.4b are MAR.
- (b) (U, X, W) have a $(p+2)$ -variate normal distribution.
- (c) Nondifferential measurement error (NDME): $\beta_{uw\cdot wx} = 0$, where $\beta_{uw\cdot wx}$ are the p regression coefficients of W in the $(p$ -variate) regression of U on W and X .

The NDME assumption is valid when the measurement error in W is unrelated to values of U , given the true values of X . As discussed in the following, this assumption allows us to estimate the p partial covariances $\sigma_{ux\cdot w}$ of U and X given W , for which there are otherwise no unique ML estimates from the external calibration pattern.

Specifically, factor the joint normal distribution of (U, X, W) as follows:

$$W \sim N(\mu_w, \sigma_{ww}); \quad (U, X | W) \sim N\left(\begin{pmatrix} \beta_{uw\cdot w} \\ \beta_{xw\cdot w} \end{pmatrix}, \begin{pmatrix} \sigma_{uu\cdot w} & \sigma_{ux\cdot w} \\ \sigma_{ux\cdot w} & \sigma_{xx\cdot w} \end{pmatrix}\right).$$

The likelihood factors into (i) the likelihood for the parameters (μ_w, σ_{ww}) of the marginal normal distribution of W based on all the data, (ii) the likelihood for the parameters $(\beta_{uw\cdot w}, \sigma_{uu\cdot w})$ of the conditional normal distribution of U given W based on the main sample, and (iii) the likelihood for the parameters $(\beta_{xw\cdot w}, \sigma_{xx\cdot w})$ of the conditional normal distribution of X given W based on the calibration sample. Hence assuming distinct parameters, by the factored likelihood method, (i) ML estimates $(\hat{\mu}_w, \hat{\sigma}_{ww})$ of (μ_w, σ_{ww}) are the sample mean and variance of W from the combined main and calibration samples; (ii) ML estimates $(\hat{\beta}_{uw\cdot w}, \hat{\sigma}_{uu\cdot w})$ are the least-squares estimates and residual covariance matrix from the regression of U on W , from the main sample; and (iii) ML estimates $(\hat{\beta}_{xw\cdot w}, \hat{\sigma}_{xx\cdot w})$ are the least-squares estimates and residual variance from the regression of X on W from the calibration sample.

By properties of the multivariate normal distribution, the NDME assumption implies that

$$\beta_{uw\cdot wx} = \beta_{uw\cdot w} - \frac{\sigma_{ux\cdot w}\beta_{xw\cdot w}}{\sigma_{xx\cdot w}} = 0, \quad \text{or} \quad \sigma_{ux\cdot w} = \frac{\beta_{uw\cdot w}\sigma_{xx\cdot w}}{\beta_{xw\cdot w}}.$$

Hence, the ML estimates of the remaining parameters are

$$\hat{\sigma}_{ux\cdot w} = \frac{\hat{\beta}_{uw\cdot w}\hat{\sigma}_{xx\cdot w}}{\hat{\beta}_{xw\cdot w}}.$$

Finally, the ML estimates of the parameters of the regression of Y on X and Z can be computed by sweep operations, reverse sweeping on W to make

W a dependent variable, and sweeping on X and Z to make those variables predictors.

This completes the description of the ML algorithm, except for one minor caveat. The estimate of the residual variance of X given (Y, Z, U) could be negative, given the fact that estimates are being combined from two independent samples. If this happens, the estimated residual variance should be set to zero. This is unlikely to happen unless X and W are weakly correlated, in which case the calibration data have limited utility.

The distinctness condition for factoring the likelihood is satisfied here because the model is “just-identified,” in the sense that the number of parameters p about which there is no information in the data (namely, the partial covariances σ_{ux-w}) is the same as the number of restrictions on the parameters from the NDME assumption.

This example considers ML estimation – a useful extension is to create multiple imputations of the missing values in Figure 1.4b, as discussed in Guo et al. (2011).

Problems

- 7.1 Assume the data in Example 7.1 are missing always at random (MAAR). Show that given (y_{11}, \dots, y_{n1}) , $\hat{\beta}_{20:1}$ and $\hat{\beta}_{21:1}$ are unbiased for $\beta_{20:1}$ and $\beta_{21:1}$. Hence, show that $\hat{\mu}_2$ is unbiased for μ_2 , assuming appropriate random sampling.
- 7.2 Assume the data in Example 7.1 are missing always completely at random (MACAR). By first conditioning on (y_{11}, \dots, y_{n1}) , find the exact small-sample variance of $\hat{\mu}_2$. (*Hint:* If u is chi-squared on d degrees of freedom, then $E(1/u) = 1/(d - 2)$) (see Morrison 1971). Hence show that $\hat{\mu}_2$ has a smaller sampling variance than \bar{y}_2 if and only if $\rho^2 > 1/(r - 2)$, where r is the number of completely observed units.
- 7.3 Compare the asymptotic variance of $\hat{\mu}_2 - \mu_2$ given by (7.13) and (7.14) with the small-sample variance computed in Problem 7.2.
- 7.4 Prove the six results on Bayes’ inference for monotone bivariate normal data in Section 7.3 (For help, see chapter 2 of Box and Tiao (1973) or chapter 18 of Gelman et al. (2013); also see the material in Section 6.1.4.)
- 7.5 For the bivariate normal distribution, express the regression coefficient $\beta_{12:2}$ of Y_1 on Y_2 in terms of the parameters ϕ in Section 7.2 and hence, derive its ML estimate for the data in Example 7.2.

7.6 Compute the large-sample variance of $\hat{\beta}_{12\cdot 2}$ in Problem 7.5 and compare with the variance of the complete-case estimate, assuming MACAR.

7.7 Show that for the setup of Problem 7.6, the estimate of $\beta_{12\cdot 2}$ obtained by maximizing the complete data loglikelihood over parameters and missing data is $\tilde{\beta}_{12\cdot 2} = \hat{\beta}_{12\cdot 2}\hat{\sigma}_{22}^*/\hat{\sigma}_{22}^*$, where (in the notation of Section 7.2),

$$\hat{\sigma}_{22}^* = \hat{\beta}_{21\cdot 1}^2 \hat{\sigma}_{11} + n^{-1} \sum_{i=1}^r (y_{i2} - \bar{y}_2 - \hat{\beta}_{21\cdot 1}(y_{i1} - \bar{y}_1))^2.$$

Hence, show that $\tilde{\beta}_{12\cdot 2}$ is not consistent for $\beta_{12\cdot 2}$ unless the fraction of missing data tends to zero as $n \rightarrow \infty$ (see Section 6.3; for help see Little and Rubin 1983b).

7.8 Show that the factored likelihood of Example 7.1 does not yield distinct parameters $\{\phi_j\}$ for a bivariate normal sample with means (μ_1, μ_2) , correlation ρ , and *common* variance σ^2 , with missing values on Y_2 .

7.9 By simulation, generate a bivariate normal sample of 20 units with parameters $\mu_1 = \mu_2 = 0$, $\sigma_{11} = \sigma_{12} = 1$, and $\sigma_{22} = 2$, and delete values of Y_2 so that $\Pr(m_{i2} = 1 | y_{i1}, y_{i2})$ equals 0.2 if $y_{i1} < 0$ and 0.8 if $y_{i1} \geq 0$.

- (a) Construct a test for whether the data are MACAR and implement the test on your data set.
- (b) Compute 95% confidence intervals for μ_2 using (i) the data before values were deleted; (ii) the complete units; and (iii) the *t*-approximation in (2) of Table 7.2. Summarize the properties of these intervals for this missingness mechanism.

7.10 Prove that SWP is commutative and conclude that the order in which a set of sweeps is taken is irrelevant algebraically.

7.11 Show that RSW is the inverse operation to SWP.

7.12 Show how to compute partial correlations and multiple correlations using SWP.

7.13 Estimate the parameters of the distribution of X_1 , X_2 , X_3 , and X_5 in Example 7.8, pretending X_4 is never observed. Would the calculations be more or less work if X_3 rather than X_4 were never observed?

7.14 Create a factorization table (see Rubin 1974) for the data in Example 7.11. State why the estimates produced in Example 7.11 are ML.

- 7.15** If data are MAR and the data analyst discards values to yield a data set with all complete-data factors, are the resultant missing data necessarily MAR? Provide an example to illustrate important points.
- 7.16**
- (i) Consider the following simple form of the discriminant analysis model for bivariate data with binary X and continuous Y :
 - (a) X is Bernoulli with $\Pr(X = 1 | \pi) = 1 - \Pr(X = 0 | \pi) = \pi$ and
 - (b) Y given $X = j$ is normal with mean μ_j and variance σ^2 ($j = 1, 0$). Derive ML estimates of $(\pi, \mu_0, \mu_1, \sigma^2)$ and the *marginal* mean and variance of Y for a complete random sample (x_i, y_i) , $i = 1, \dots, n$ on X and Y .
 - (ii) Suppose now that X is completely observed, but $n - r$ values of Y are missing, with an ignorable mechanism. Use the methods of Chapter 7 to derive the ML estimates of the marginal mean and variance of Y for this monotone pattern.
 - (iii) Describe how to generate draws from the posterior distribution of the parameters $(\pi, \mu_0, \mu_1, \sigma^2)$, when the prior distribution takes the form $p(\pi, \mu_0, \mu_1, \log \sigma^2) \propto \pi^{1/2}(1 - \pi)^{1/2}$.
- 7.17** For the model of Problem 7.16, consider now the reverse monotone missing data pattern, with Y completely observed but $n - r$ values of X missing, and an ignorable mechanism. Does the factored likelihood method provide closed-form expressions for ML estimates for this pattern? (*Hint:* Find the conditional distribution of X given Y and the marginal distribution of Y . Are the parameters of these two distributions distinct?)
- 7.18** Outline extensions of Problem 7.16 to multivariate monotone patterns where the factored likelihood method works.

8

Maximum Likelihood for General Patterns of Missing Data: Introduction and Theory with Ignorable Nonresponse

8.1 Alternative Computational Strategies

Patterns of incomplete data in practice often do not have the particular forms that allow explicit maximum likelihood (ML) estimates to be calculated by exploiting factorizations of the likelihood. Furthermore, for some models a factorization exists, but the parameters in the factorization are not distinct, and thus maximizing the factors separately does not maximize the likelihood. In this chapter, we consider iterative methods of computation for situations without explicit ML estimates. In some cases, these methods can be *applied* to incomplete-data factors discussed in Section 7.5.

Suppose, as before, that we have a model for the complete data Y , with associated density $f(Y | \theta)$ indexed by unknown parameter θ , generally a vector. We write $Y = (Y_{(0)}, Y_{(1)})$, where $Y_{(0)}$ represents the observed part of Y and $Y_{(1)}$ denotes the missing part. In this chapter, we assume for simplicity that the data are missing at random (MAR) and that the objective is to maximize the likelihood

$$L(\theta | Y_{(0)}) = \int f(Y_{(0)}, Y_{(1)} | \theta) dY_{(1)} \quad (8.1)$$

with respect to θ . Similar considerations apply to the more general case when the data are not MAR, and consequently a factor representing the missingness mechanism is included in the model; such cases are considered in Chapter 15.

When the likelihood is differentiable and unimodal, ML estimates can be found by solving the likelihood equation

$$D_\ell(\theta | Y_{(0)}) \equiv \frac{\partial \ln L(\theta | Y_{(0)})}{\partial \theta} = 0. \quad (8.2)$$

When a closed-form solution of (8.2) cannot be found, iterative methods can be applied. Let $\theta^{(0)}$ be an initial estimate of θ in its parameter space, perhaps, for

example, an estimate based on the completely observed units, or the complete-data estimate of θ after the missing data $Y_{(1)}$ have been filled in by some approximations. Let $\theta^{(t)}$ be the estimate at the t th iteration. The *Newton–Raphson* algorithm is defined by the equation

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)} | Y_{(0)}) D_\ell(\theta^{(t)} | Y_{(0)}), \quad (8.3)$$

where $I(\theta | Y_{(0)})$ is the observed information, defined as follows:

$$I(\theta | Y_{(0)}) = -\frac{\partial^2 \ell(\theta | Y_{(0)})}{\partial \theta \partial \theta}.$$

If the loglikelihood function is concave and unimodal, then the sequence of iterates $\theta^{(t)}$ converges to the ML estimate $\hat{\theta}$ of θ , in one step if the loglikelihood is a quadratic function of θ . A variant of this procedure is the Method of Scoring, where the observed information in (8.3) is replaced by the expected information:

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}(\theta^{(t)}) D_\ell(\theta^{(t)} | Y_{(0)}), \quad (8.4)$$

where the expected information is defined, as earlier, by

$$J(\theta) = E\{I(\theta | Y_{(0)}) | \theta\} = - \int \frac{\partial^2 \ell(\theta | Y_{(0)})}{\partial \theta \partial \theta} f(Y_{(0)} | \theta) dY_{(0)}.$$

Both of these methods involve calculating the matrix of second derivatives of the loglikelihood. For complex patterns of incomplete data, the entries in this matrix tend to be complicated functions of θ . Also, the matrix is large when θ has high dimension. As a result, to be practical, the methods can require careful algebraic manipulations and efficient programming. Other variants of the Newton–Raphson algorithm approximate the derivatives of the loglikelihood numerically, by using first and second differences between successive iterates.

An alternative computing strategy for incomplete-data problems, which does not require second derivatives to be calculated or approximated, is the *Expectation–Maximization (EM)* algorithm, a method that relates ML estimation of θ from $\ell(\theta | Y_{(0)})$ to ML estimation based on the complete-data loglikelihood $\ell(\theta | Y)$. Sections 8.2–8.4 of this chapter, as well as parts of other chapters, are devoted to the EM algorithm and its extensions.

In many important cases, the EM algorithm is remarkably simple, both conceptually and computationally. However, EM does *not* share with Newton–Raphson or scoring algorithms the property of yielding estimates asymptotically equivalent to ML estimates after a single iteration, and thus EM generally requires repeated iterations. Also, EM can be very slow to converge; Dempster et al. (1977) show that each iteration of EM increases the likelihood, and convergence is linear with rate of increase proportional to the fraction of information

about θ in $\ell(\theta | Y)$ that is observed, in a sense made precise in Section 8.4.3. Furthermore, in some problems, the M step is difficult, e.g., has no closed form, and then the theoretical simplicity of EM does not convert to practical simplicity. However, there are two types of extensions of EM that often can avoid these drawbacks.

The first type of extension retains the simplicity of implementation, relying on complete-data computations. These algorithms retain EM's monotone increase in the likelihood and its stable convergence to a local maximum. Because these algorithms are so similar to EM, we call them generically "EM-type" algorithms. The EM-type algorithms described here include ECM (Section 8.5.1), ECME (Section 8.5.2), AECM (Section 8.5.2), and PX-EM (Section 8.5.3). ECM replaces the M step of EM with two or more conditional (on parameters) maximization steps. ECME is a variant of ECM in which the CM steps maximize either the usual complete-data log likelihood or the actual log-likelihood. AECM is an extension of ECME that allows alternative CM steps to maximize different complete-data loglikelihoods, corresponding to different definitions of missing data. PX-EM is a bigger change in that it expands the parameter space over which the maximization is taking place to include parameters whose values are known, thereby often greatly speeding EM. A related idea is that of efficient augmentation, where the missing-data structure is optimized in advance to speed the resultant EM.

The second type of EM extension mixes EM with other techniques, which can result in efficient algorithms but typically without the guaranteed monotone increase in the likelihood. Versions of the second type of extension of EM are discussed in Section 8.6. They include switching from EM to a Newton–Raphson method after some initial EM iterations, the gradient EM algorithm of Lange (1995a), and the accelerated EM method of Jamshidian and Jennrich (1993), which is based on generalized conjugate-gradient ideas. McLachlan and Krishnan (1997) provide an excellent review of the EM algorithm and these extensions, including more theoretical results and details than in this volume. We focus on missing-data applications.

8.2 Introduction to the EM Algorithm

The EM algorithm is a general iterative algorithm for ML estimation in incomplete-data problems. In fact, the range of problems that can be attacked by EM is remarkably broad (Meng and Pedlow 1992), and includes ML for problems not usually considered to involve missing data, such as variance-component estimation and factor analysis (see also Becker et al. 1997).

The EM algorithm formalizes a relatively old ad hoc idea for handling missing data, already introduced in Chapter 2: (i) replace missing values by estimated values, (ii) estimate parameters, (iii) re-estimate the missing values assuming

the new parameter estimates are correct, (iv) re-estimate parameters, and so forth, iterating until apparent convergence. Such methods are EM algorithms for models where the complete-data loglikelihood, $\ell(\theta | Y_{(0)}, Y_{(1)}) = \ln L(\theta | Y_{(0)}, Y_{(1)})$, is linear in $Y_{(1)}$; more generally, missing sufficient statistics rather than individual missing values need to be estimated, in step (iii) above and even more generally, the loglikelihood $\ell(\theta | Y)$ itself needs to be estimated at each iteration of the algorithm.

Because the EM algorithm is so closely tied to the intuitive idea of filling in missing values and iterating, it is not surprising that the algorithm has been proposed for many years in special contexts. The earliest reference seems to be McKendrick (1926), which considers it in a medical application. Hartley (1958) considers the general case of counted data and develops the theory quite extensively; many of the key ideas can be found there. Baum et al. (1970) use the algorithm in a Markov model, and they prove key mathematical results in this case that generalize quite easily. Orchard and Woodbury (1972) first notes the general applicability of the underlying idea, calling it the “missing information principle.” Sundberg (1974) explicitly considers properties of the general likelihood equations, and Beale and Little (1975) further develops the theory for the normal model. The term “EM” was introduced in Dempster et al. (1977), and this work exposed the full generality of the algorithm by (i) proving general results about its behavior, specifically that each iteration increases the likelihood $\ell(\theta | Y_{(0)})$, and (ii) providing a wide range of examples. Since 1977, there have been many new uses of the EM algorithm, as well as further work on its convergence properties (Wu 1983).

Each iteration of EM consists of an expectation step (E step) and a maximization step (M step). These steps are often easy to construct conceptually, to program for calculation, and to fit into computer storage. Each step also has a direct statistical interpretation. An additional advantage of the algorithm is that it can be shown to converge reliably, in the sense that under general conditions, each iteration increases the loglikelihood $\ell(\theta | Y_{(0)})$, and if $\ell(\theta | Y_{(0)})$ is bounded, the sequence $\{\ell(\theta^{(t)} | Y_{(0)}), t = 0, 1, 2, \dots\}$ converges to a stationary value of $\ell(\theta | Y_{(0)})$. Quite generally, if the sequence $\{\theta^{(t)}, t = 0, 1, 2, \dots\}$ converges, it converges to a local maximum or saddle point of $\ell(\theta | Y_{(0)})$.

8.3 The E Step and The M Step of EM

The M step is particularly simple to describe: perform ML estimation of θ just as if there were no missing data, that is, as if they had been filled in. Thus the M step of EM uses the identical computational method as ML estimation from $\ell(\theta | Y)$ with Y completely observed.

The E step finds the conditional expectation of the “missing data” given the observed data and current estimated parameters, and then substitutes these expectations for the “missing data.” The term “missing data” is written with quotes because EM does not necessarily substitute the missing values themselves. The key idea of EM, which delineates it from the ad hoc idea of filling in missing values and iterating, is that “missing data” are generally not $Y_{(0)}$ but the functions of $Y_{(0)}$ appearing in the complete-data loglikelihood $\ell(\theta | Y)$.

Specifically, let $\theta^{(t)}$ be the current estimate of the parameter θ . The E step of EM finds the expected complete-data loglikelihood if θ were $\theta^{(t)}$:

$$Q(\theta | \theta^{(t)}) = \int \ell(\theta | Y) f(Y_{(0)} | Y_{(1)}, \theta = \theta^{(t)}) dY_{(0)}.$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected complete-data loglikelihood:

$$Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}), \quad \text{for all } \theta.$$

Example 8.1 Univariate Normal Data. Suppose y_i are iid $N(\mu, \sigma^2)$ where y_i for $i = 1, \dots, r$ are observed, and y_i for $i = r+1, \dots, n$ are missing, and assume the missingness is ignorable. The expectation of each missing y_i given Y_{obs} and $\theta = (\mu, \sigma^2)$ is μ .

However, from Example 6.1, the loglikelihood $\ell(\theta | Y)$, based on all y_i , $i = 1, \dots, n$, is linear in the sufficient statistics $\sum_1^n y_i$ and $\sum_1^n y_i^2$. Thus, the E step of the algorithm calculates

$$E\left(\sum_{i=1}^n y_i | \theta^{(t)}, Y_{(0)}\right) = \sum_{i=1}^r y_i + (n-r)\mu^{(t)} \tag{8.5}$$

and

$$E\left(\sum_{i=1}^n y_i^2 | \theta^{(t)}, Y_{(0)}\right) = \sum_{i=1}^r y_i^2 + (n-r)[(\mu^{(t)})^2 + (\sigma^{(t)})^2], \tag{8.6}$$

for current estimates $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$ of the parameters. Note that simply substituting $\mu^{(t)}$ for the missing values y_{r+1}, \dots, y_n is not correct, because it leads to the omission of the term $(n-r)(\sigma^{(t)})^2$ in (8.6).

With no missing data, the ML estimate of μ is $\sum_{i=1}^n y_i/n$ and the ML estimate of σ^2 is $\sum_{i=1}^n y_i^2/n - (\sum_{i=1}^n y_i/n)^2$. The M step uses these same expressions with

the current expectations of the sufficient statistics calculated in the E step substituted for the incompletely observed (and hence missing) sufficient statistics. Thus the M step calculates

$$\mu^{(t+1)} = E \left(\sum_{i=1}^n y_i \mid \theta^{(t)}, Y_{(0)} \right) / n, \quad (8.7)$$

$$(\sigma^{(t+1)})^2 = E \left(\sum_{i=1}^n y_i^2 \mid \theta^{(t)}, Y_{(0)} \right) / n - (\mu^{(t+1)})^2. \quad (8.8)$$

Setting $\mu^{(t)} = \mu^{(t+1)} = \hat{\mu}$ and $\sigma^{(t)} = \sigma^{(t+1)} = \hat{\sigma}$ in Eqs. (8.5)–(8.8) shows that a fixed point of these iterations is

$$\hat{\mu} = \sum_{i=1}^r y_i / r$$

and

$$\hat{\sigma}^2 = \sum_{i=1}^r y_i^2 / r - \hat{\mu}^2,$$

which are the ML estimates of μ and σ^2 from $Y_{(0)}$ assuming MAR and distinctness of parameters. Of course, the EM algorithm is unnecessary for this example, because the explicit ML estimates ($\hat{\mu}, \hat{\sigma}^2$) are available.

Example 8.2 A Multinomial Example (Dempster et al. 1977). Suppose that the data vector of observed counts $Y_{(0)} = (38, 34, 125)$ is postulated to arise from a multinomial with cell probabilities $(1/2 - \theta/2, \theta/4, 1/2 + \theta/4)$. The objective is to find the ML estimate of θ . Define $Y = (y_1, y_2, y_3, y_4)$ to be multinomial with probabilities $(1/2 - \theta/2, \theta/4, \theta/4, 1/2)$, where $Y_{(0)} = (y_1, y_2, y_3 + y_4)$. Notice that if Y were observed, the ML estimate of θ would be immediate:

$$\hat{\theta}_{\text{complete}} = \frac{y_2 + y_3}{y_1 + y_2 + y_3}. \quad (8.9)$$

Also note that the loglikelihood $\ell(\theta \mid Y)$ is linear in Y , so finding the expectation of $\ell(\theta \mid Y)$ given θ and $Y_{(0)}$ involves finding the expectation of Y given θ and $Y_{(0)}$, which in effect fills in estimates of the missing values:

$$\begin{aligned} E(y_1 \mid \theta, Y_{(0)}) &= 38, \\ E(y_2 \mid \theta, Y_{(0)}) &= 34, \\ E(y_3 \mid \theta, Y_{(0)}) &= 125(\theta/4)/(1/2 + \theta/4), \\ E(y_4 \mid \theta, Y_{(0)}) &= 125(1/2)/(1/2 + \theta/4). \end{aligned}$$

Table 8.1 The EM algorithm for Example 8.2

t	$\theta^{(t)}$	$\theta^{(t)} - \hat{\theta}$	$(\theta^{(t+1)} - \hat{\theta}) / (\theta^{(t)} - \hat{\theta})$
0	0.500 000 000	0.126 821 498	0.146 5
1	0.608 247 423	0.018 574 075	0.134 6
2	0.624 321 051	0.002 500 447	0.133 0
3	0.626 488 879	0.000 332 619	0.132 8
4	0.626 777 323	0.000 044 176	0.132 8
5	0.626 815 632	0.000 005 866	0.132 8
6	0.626 820 719	0.000 000 779	
7	0.626 821 395	0.000 000 104	
8	0.626 821 484	0.000 000 014	

Thus at the t th iteration, with estimate $\theta^{(t)}$, we have for the E step

$$y_3^{(t)} = 125(\theta^{(t)}/4)/(1/2 + \theta^{(t)}/4), \quad (8.10)$$

and for the M step, from (8.9) we have

$$\theta^{(t+1)} = (34 + y_3^{(t)}) / (72 + y_3^{(t)}). \quad (8.11)$$

Iterating between (8.10) and (8.11) defines the EM algorithm for this problem. In fact, setting $\theta^{(t+1)} = \theta^{(t)} = \hat{\theta}$ and combining the two equations yields a quadratic equation in $\hat{\theta}$ and thus a closed-form solution for the ML estimate. Table 8.1 shows the linear convergence of EM to this solution starting from $\theta^{(0)} = 1/2$; many decimal points are shown to allow calculation of the rate of convergence in the last column.

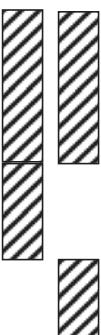
Example 8.3 Bivariate Normal Sample with Missingness on Both Variables. A simple but nontrivial application of EM is for a bivariate normal sample with a general pattern of missingness: The first group of units has both Y_1 and Y_2 observed, the second group has Y_1 observed and Y_2 missing, and the third group has Y_2 observed and Y_1 missing (see Figure 8.1). We wish to calculate the ML estimates of μ and Σ , the mean vector and covariance matrix of Y_1 and Y_2 .

Like Example 8.1, but unlike Example 8.2, filling in missing values in the E step does not work because the loglikelihood $\ell(\theta | y)$ is not linear in the data, but rather is linear in the following sufficient statistics:

$$s_1 = \sum_1^n y_{i1}, \quad s_2 = \sum_1^n y_{i2}, \quad s_{11} = \sum_1^n y_{i1}^2, \quad s_{22} = \sum_1^n y_{i2}^2, \quad s_{12} = \sum_1^n y_{i1}y_{i2}, \quad (8.12)$$

Y_1 Y_2

Figure 8.1 The missingness pattern for Example 8.3.



which are simple functions of the sample means, variances, and covariances. The task at the E step is thus to find the conditional expectation of the sums in (8.12), given $Y_{(0)}$ and $\theta = (\mu, \Sigma)$. For the group of units with both y_{i1} and y_{i2} observed, the conditional expectations of the quantities in (8.12) equal their observed values. For the group of units with y_{i1} observed but y_{i2} missing, the expectations of y_{i1} and y_{i1}^2 equal their observed values; the expectations of y_{i2} , y_{i2}^2 and $y_{i1}y_{i2}$ are found from the regression of the y_{i2} on y_{i1} :

$$\begin{aligned} E(y_{i2} | y_{i1}, \mu, \Sigma) &= \beta_{20 \cdot 1} + \beta_{21 \cdot 1}y_{i1}, \\ E(y_{i2}^2 | y_{i1}, \mu, \Sigma) &= (\beta_{20 \cdot 1} + \beta_{21 \cdot 1}y_{i1})^2 + \sigma_{22 \cdot 1}, \\ E(y_{i2}y_{i1} | y_{i1}, \mu, \Sigma) &= (\beta_{20 \cdot 1} + \beta_{21 \cdot 1}y_{i1})y_{i1}, \end{aligned}$$

where $\beta_{20 \cdot 1}$, $\beta_{21 \cdot 1}$, and $\sigma_{22 \cdot 1}$ are functions of Σ corresponding to the regression of y_{i2} on y_{i1} (see Example 7.1 for details). For the units with y_{i2} observed and y_{i1} missing, the regression of y_{i1} on y_{i2} is used to calculate the missing contributions to the sufficient statistics. Having found the expectations of y_{i1} , y_{i2} , y_{i1}^2 , y_{i2}^2 , and $y_{i1}y_{i2}$ for each unit in the three groups, the expectations of the sufficient statistics in (8.12) are found as the sums of these quantities over all n units. The M step calculates the usual moment-based ML estimates of μ and Σ from those filled-in sufficient statistics:

$$\begin{aligned} \hat{\mu}_1 &= s_1/n, \quad \hat{\mu}_2 = s_2/n, \\ \hat{\sigma}_1^2 &= s_{11}/n - \hat{\mu}_1^2, \quad \hat{\sigma}_2^2 = s_{22}/n - \hat{\mu}_2^2, \quad \hat{\sigma}_{12} = s_{12}/n - \hat{\mu}_1\hat{\mu}_2. \end{aligned}$$

The EM algorithm then carries out these steps iteratively. This algorithm is generalized in Chapter 9 to a multivariate normal distribution with any pattern of missing values.

In the above example, the mean μ and covariance matrix Σ were unrestricted, aside from the constraint that Σ is positive semidefinite. Often, we are interested in computing ML estimates for models that place constraints on parameters. For example, in normal models for repeated measures, we might constrain the covariance matrix to have a compound-symmetry structure, or in loglinear models for contingency tables, we may fit models that constrain the cell probabilities, assuming that particular higher-order associations are zero. When EM is applied to fit models with parameter constraints, a useful feature is that parameter constraints do not affect the E step, which is the missing-data part of the problem. The M step maximizes the expected complete-data loglikelihood over parameters, *subject to parameter constraints*, given current estimates of complete-data sufficient statistics. If this step yields explicit estimates with complete data, this is an easy modification of EM for the unconstrained model, and in other cases standard software may be available. Examples of this attractive property of EM are given in Examples 11.1–11.3.

8.4 Theory of the EM Algorithm

8.4.1 Convergence Properties of EM

The distribution of the complete data Y can be factored as follows:

$$\ell(Y \mid \theta) = \ell(Y_{(0)}, Y_{(1)} \mid \theta) = \ell(Y_{(0)} \mid \theta) \ell(Y_{(1)} \mid Y_{(0)}, \theta),$$

where $\ell(Y_{(0)} \mid \theta)$ is the density of the observed data $Y_{(0)}$ and $\ell(Y_{(1)} \mid Y_{(0)}, \theta)$ is the density of the missing data given the observed data. The corresponding decomposition of the loglikelihood is

$$\ell(\theta \mid Y) = \ell(\theta \mid Y_{(0)}, Y_{(1)}) = \ell(\theta \mid Y_{(0)}) + \text{lnf}(Y_{(1)} \mid Y_{(0)}, \theta).$$

The objective is to estimate θ by maximizing the incomplete-data loglikelihood $\ell(\theta \mid Y_{(0)})$ with respect to θ for fixed observed $Y_{(0)}$; this task, however, can be difficult to accomplish directly.

First, write

$$\ell(\theta \mid Y_{(0)}) = \ell(\theta \mid Y) - \text{lnf}(Y_{(1)} \mid Y_{(0)}, \theta), \quad (8.13)$$

where $\ell(\theta \mid Y_{(0)})$ is the observed loglikelihood to be maximized, $\ell(\theta \mid Y)$ is the complete-data loglikelihood, which we assume is relatively easy to maximize, and $\text{lnf}(Y_{(1)} \mid Y_{(0)}, \theta)$ can be viewed as the missing part of the complete-data loglikelihood.

The expectation of both sides of (8.13) over the distribution of the missing data $Y_{(1)}$, given the observed data $Y_{(0)}$ and a current estimate of θ , say $\theta^{(t)}$, is

$$\ell(\theta \mid Y_{(0)}) = Q(\theta \mid \theta^{(t)}) - H(\theta \mid \theta^{(t)}), \quad (8.14)$$

where

$$Q(\theta \mid \theta^{(t)}) = \int [\ell(\theta \mid Y_{(0)}, Y_{(1)})] f(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}) dY_{(1)} \quad (8.15)$$

and

$$H(\theta \mid \theta^{(t)}) = \int [\ln f(Y_{(1)} \mid Y_{(0)}, \theta)] f(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}) dY_{(1)}. \quad (8.16)$$

Note that

$$H(\theta \mid \theta^{(t)}) \leq H(\theta^{(t)} \mid \theta^{(t)}) \quad (8.17)$$

by Jensen's inequality (see Rao 1972, p. 47).

Consider a sequence of iterates $(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots)$. The difference in values of $\ell(\theta \mid Y_{(0)})$ at successive iterates is given by

$$\begin{aligned} \ell(\theta^{(t+1)} \mid Y_{(0)}) - \ell(\theta^{(t)} \mid Y_{(0)}) &= [Q(\theta^{(t+1)} \mid \theta^{(t)}) - Q(\theta^{(t)} \mid \theta^{(t)})] \\ &\quad - [H(\theta^{(t+1)} \mid \theta^{(t)}) - H(\theta^{(t)} \mid \theta^{(t)})]. \end{aligned} \quad (8.18)$$

An EM algorithm chooses $\theta^{(t+1)}$ to maximize $Q(\theta \mid \theta^{(t)})$ with respect to θ . More generally, a GEM (generalized expectation–maximization) algorithm chooses $\theta^{(t+1)}$ so that $Q(\theta^{(t+1)} \mid \theta^{(t)})$ is greater than $Q(\theta^{(t)} \mid \theta^{(t)})$. Hence, the difference of Q functions in (8.18) is positive for any EM or GEM algorithm. Furthermore, note that the difference in the H functions in (8.18) is negative by (8.17). Hence, for any EM or GEM algorithm, the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ increases the loglikelihood. This proves the following theorem, which is a key mathematical result of Dempster et al. (1977).

Theorem 8.1 *Every GEM algorithm increases $\ell(\theta \mid Y_{(0)})$ at each iteration, that is,*

$$\ell(\theta^{(t+1)} \mid Y_{(0)}) \geq \ell(\theta^{(t)} \mid Y_{(0)}),$$

with equality if and only if

$$Q(\theta^{(t+1)} \mid \theta^{(t)}) = Q(\theta^{(t)} \mid \theta^{(t)}).$$

Corollary 8.1 *Define $M(\cdot)$ such that $\theta^{(t+1)} = M(\theta^{(t)})$. Suppose that for some θ^* in the parameter space of θ , $\ell(\theta^* \mid Y_{(0)}) \geq \ell(\theta \mid Y_{(0)})$ for all θ . Then for every GEM algorithm,*

$$\ell(M(\theta^*) \mid Y_{(0)}) = \ell(\theta^* \mid Y_{(0)}),$$

$$Q(M(\theta^*) \mid \theta^*) = Q(\theta^* \mid \theta^*),$$

and

$$f(Y_{(1)} \mid Y_{(0)}, M(\theta^*)) = f(Y_{(1)} \mid Y_{(0)}, \theta^*).$$

Corollary 8.2 Suppose that for some θ^* in the parameter space of θ $\ell(\theta^* \mid Y_{(0)}) > \ell(\theta \mid Y_{(0)})$ for all θ . Then for every GEM algorithm:

$$M(\theta^*) = \theta^*.$$

Theorem 8.1 implies that $\ell(\theta \mid Y_{(0)})$ is nondecreasing after each iteration of a GEM algorithm, and is strictly increasing after any iteration such that Q increases; that is, whenever $Q(\theta^{(t+1)} \mid \theta^{(t)}, Y_{(0)}) > Q(\theta^{(t)} \mid \theta^{(t)}, Y_{(0)})$. The corollaries imply that a ML estimate of θ is a fixed point of a GEM algorithm.

Another important result concerning EM algorithms is given by Theorem 8.2, which applies when $Q(\theta \mid \theta^{(t)})$ is maximized by setting the first derivative equal to zero.

Theorem 8.2 Suppose a sequence of EM iterates is such that

- (a) $D^{10}Q(\theta^{(t+1)} \mid \theta^{(t)}) = 0$, where “D” here denotes derivative, and D^{10} means the derivative with respect to the first argument, that is, define

$$D^{10}Q(\theta^{(t+1)} \mid \theta^{(t)}) = \frac{\partial}{\partial \theta} Q(\theta \mid \theta^{(t)}) \Big|_{\theta=\theta^{(t+1)}},$$

(b) $\theta^{(t)}$ converges to θ^* , and

(c) $f(Y_{(1)} \mid Y_{(0)}, \theta)$ is smooth in θ , where smooth is defined in the proof.

Then

$$D\ell(\theta^* \mid Y_{(0)}) \equiv \frac{\partial}{\partial \theta} \ell(\theta \mid Y_{(0)}) \Big|_{\theta=\theta^*} = 0$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof:

$$\begin{aligned} D\ell(\theta^{(t+1)} \mid Y_{(0)}) &= D^{10}Q(\theta^{(t+1)} \mid \theta^{(t)}) - D^{10}H(\theta^{(t+1)} \mid \theta^{(t)}) \\ &= -D^{10}H(\theta^{(t+1)} \mid \theta^{(t)}) \\ &= -\frac{\partial}{\partial \theta} \int [\ln f(Y_{(1)} \mid Y_{(0)}, \theta)] f(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}) dY_{(1)} \Big|_{\theta=\theta^{(t+1)}}, \end{aligned}$$

which, assuming sufficient smoothness to interchange the order of differentiation and integration, converges to

$$-\int \frac{\partial}{\partial \theta} f(Y_{(1)} | Y_{(0)}, \theta) dY_{(1)} \Big|_{\theta=\theta^{(t+1)}},$$

which equals zero after again interchanging the order of integration and differentiation. \square

Other EM results in Dempster et al. (1977) and Wu (1983) regarding convergence include the following:

1. If $f(Y | \theta)$ is a general (curved) exponential family and $\ell(\theta | Y_{(0)})$ is bounded, then $\ell(\theta^{(t)} | Y_{(0)})$ converges to a stationary value ℓ^* .
2. If $f(Y | \theta)$ is a regular exponential family and $\ell(\theta | Y_{(0)})$ is bounded, then $\theta^{(t)}$ converges to a stationary point θ^* .
3. If $\ell(\theta | Y_{(0)})$ is bounded, $\ell(\theta^{(t)} | Y_{(0)})$ converges to some ℓ^* .

Points 2 and 3 concern EM for exponential families, which is the topic of the next subsection.

8.4.2 EM for Exponential Families

The EM algorithm has a particularly simple and useful interpretation when the complete data Y have a distribution from the *regular exponential family* defined by

$$f(Y | \theta) = b(Y)\exp(s(Y)\theta - a(\theta)), \quad (8.19)$$

where θ denotes a $(d \times 1)$ parameter vector, $s(Y)$ denotes a $(1 \times d)$ vector of *complete-data* sufficient statistics, and a and b are functions of θ and Y , respectively. Many complete-data problems can be modeled by a distribution of the form (8.19), which includes as special cases essentially all the examples in Parts II and III of this book. The E step for iteration $(t + 1)$ given (8.19) involves estimating the complete-data sufficient statistics $s(Y)$ by

$$s^{(t+1)} = E(s(Y) | Y_{(0)}, \theta^{(t)}). \quad (8.20)$$

The M step determines the new estimate $\theta^{(t+1)}$ of θ as the solution of the likelihood equations

$$E(s(Y) | \theta) = s^{(t+1)}, \quad (8.21)$$

which are simply the likelihood equations for the complete data Y with $s(Y)$ replaced by $s^{(t+1)}$. Equation (8.21) can often be solved for θ explicitly, or at least through existing complete-data computer programs. In such cases, the computational problem is effectively confined to the E step, which involves estimating (or “imputing”) the statistics $s(Y)$, using (8.20).

The following example, described in Dempster et al. (1977, 1980), applies EM in a situation where the observed data are complete but do not belong to the exponential family (8.19). ML estimation is achieved by imbedding the observed data in a larger data set belonging to the regular exponential family (8.19), and then applying EM to this augmented dataset.

Example 8.4 *ML Estimation for a Sample from the Univariate t Distribution with Known Degrees of Freedom.* Suppose that the observed data $Y_{(0)}$ consist of a random sample $X = (x_1, x_2, \dots, x_n)$ from a Student's t distribution with center μ , scale parameter σ and known degrees of freedom (df) v , with density

$$f(x_i | \theta) = \frac{\Gamma(v/2 + 1/2)}{(\pi v \sigma^2)^{1/2} \Gamma(v/2) (1 + (x_i - \mu)^2 / (v \sigma^2))^{(v+1)/2}}. \quad (8.22)$$

This distribution does not belong to the exponential family (8.19), and ML estimation of $\theta = (\mu, \sigma)$ requires an iterative algorithm. We define an augmented complete dataset $Y = (Y_{(0)}, Y_{(1)})$, where $Y_{(0)} = X$ and $Y_{(1)} = W = (w_1, w_2, \dots, w_n)^T$ is a vector of unobserved positive quantities, such that pairs (x_i, w_i) are independent across units i , with distribution specified by

$$(x_i | \theta, w_i) \sim_{\text{ind}} N(\mu, \sigma^2/w_i), \quad (w_i | \theta) \sim \chi_v^2/v, \quad (8.23)$$

where χ_v^2 denotes the chi-squared distribution with v df. Model (8.23) leads to the marginal distribution of x_i given by (8.22), so applying EM for the expanded model provides ML estimates of θ for the t model. The augmented data Y belong to the exponential family (8.19) with complete-data sufficient statistics

$$s_0 = \sum_{i=1}^n w_i, \quad s_1 = \sum_{i=1}^n w_i x_i, \quad s_2 = \sum_{i=1}^n w_i x_i^2.$$

Hence, given current parameter estimates $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$, the $(t+1)$ th iteration of EM is as follows:

E step: Compute $s_0^{(t)} = \sum_{i=1}^n w_i^{(t)}, s_1^{(t)} = \sum_{i=1}^n w_i^{(t)} x_i, s_2^{(t)} = \sum_{i=1}^n w_i^{(t)} x_i^2$ where $w_i^{(t)} = E(w_i | x_i, \theta^{(t)})$. A simple calculation shows that the distribution of w_i given (x_i, θ) is $\chi_{v+1}^2 (v + (x_i - \mu)^2 / \sigma^2)^{-1}$. Hence

$$w_i^{(t)} = E(w_i | x_i, \theta^{(t)}) = \frac{v + 1}{v + d_i^{(t)2}}, \quad (8.24)$$

where $d_i^{(t)} = (x_i - \mu^{(t)}) / \sigma^{(t)}$ is the current estimate of the number of standard deviations x_i is from $\mu^{(t)}$.

M step: Compute new estimates of θ from the estimated sufficient statistics $(s_0^{(t)}, s_1^{(t)}, s_2^{(t)})$. These are just weighted least-squares estimates:

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{s_0^{(t)}}, \quad \hat{\sigma}^{(t+1)2} = \frac{1}{n} \sum_{i=1}^n w_i^{(t)} (x_i - \hat{\mu}^{(t+1)})^2 = \frac{s_2^{(t)} - s_1^{(t)2}/s_0^{(t)}}{n}, \quad (8.25)$$

as displayed in general in Example 6.10. The EM algorithm defined by (8.24) and (8.25) is iteratively reweighted least squares, with weights $w_i^{(t)}$ that down-weight values x_i further from the mean. Thus, ML for this t model yields a form of robust estimation. Extensions of this example are given in following sections here, and in Chapter 12.

8.4.3 Rate of Convergence of EM

Note that differentiating (8.14) twice with respect to θ yields, for any $Y_{(1)}$,

$$I(\theta | Y_{(0)}) = I(\theta | Y_{(0)}, Y_{(1)}) + \partial^2 \ln f(Y_{(1)} | Y_{(0)}, \theta) / \partial \theta \partial \theta,$$

where $I(\theta | Y_{(0)}, Y_{(1)})$ is the observed information based on $Y = (Y_{(0)}, Y_{(1)})$, and the negative of the last term is the missing information from $Y_{(1)}$. Taking expectations over the distribution of $Y_{(1)}$ given $Y_{(0)}$ and θ yields

$$I(\theta | Y_{(0)}) = -D^{20}Q(\theta | \theta) + D^{20}H(\theta | \theta), \quad (8.26)$$

where Q and H are given by (8.15) and (8.16), provided differentials with respect to θ can be passed through the integral signs. If we evaluate the functions in (8.26) at the converged value θ^* of θ , and call $i_{\text{com}} = -D^{20}Q(\theta | \theta)|_{\theta=\theta^*}$ the complete information, $i_{\text{obs}} = I(\theta | Y_{(0)})|_{\theta=\theta^*}$ the observed information, and $i_{\text{mis}} = -D^{20}H(\theta | \theta)|_{\theta=\theta^*}$ the missing information, then (8.26) leads to

$$i_{\text{obs}} = i_{\text{com}} - i_{\text{mis}}, \quad (8.27)$$

which has the appealing interpretation that the observed information equals the complete information minus the missing information.

The rate of convergence of the EM algorithm is closely related to these quantities. Specifically, rearranging (8.27) and multiplying both sides by i_{com}^{-1} yields:

$$U = i_{\text{mis}} i_{\text{com}}^{-1} = I - i_{\text{obs}} i_{\text{com}}^{-1}, \quad (8.28)$$

where the matrix U measures the matrix fraction of missing information, and controls the speed of convergence of EM. Specifically, Dempster et al. (1977) show that for $\theta^{(t)}$ near θ^* ,

$$|\theta^{(t+1)} - \theta^*| = \lambda |\theta^{(t)} - \theta^*|, \quad (8.29)$$

where $\lambda = U$ for scalar θ or often the largest eigenvalue of U for vector θ ; Meng and Rubin (1994) consider situations where (8.29) is replaced by a matrix version, with λ a block diagonal matrix.

Louis (1982) rewrites the missing information in terms of complete-data quantities, and shows that

$$\begin{aligned} -D^{20}H(\theta | \theta) &= E\{D_\ell(\theta | Y_{(0)}, Y_{(1)})D_\ell^T(\theta | Y_{(0)}, Y_{(1)}) | Y_{(0)}, \theta\} \\ &\quad - D_\ell(\theta | Y_{(0)})D_\ell^T(\theta | Y_{(0)}), \end{aligned}$$

where, as earlier, D_ℓ denotes the score function. At the ML estimate $D_\ell(\hat{\theta} | Y_{(0)}) = 0$, so the last term vanishes. Equation (8.27) becomes

$$I(\hat{\theta} | Y_{(0)}) = -D^{20}Q(\hat{\theta} | \hat{\theta}) - E\{D_\ell(\theta | Y_{(0)}, Y_{(1)})D_\ell^T(\theta | Y_{(0)}, Y_{(1)}) | Y_{(0)}, \theta\}|_{\theta=\hat{\theta}}, \quad (8.30)$$

which may be useful for computations.

An expression analogous to (8.27) for the expected information $J(\theta)$ is obtained by taking expectations of (8.26) over $Y_{(0)}$. Specifically,

$$J(\theta) = J_c(\theta) + E\{D^{20}H(\theta | \theta)\}, \quad (8.31)$$

where $J_c(\theta)$ is the expected complete information based on $Y = (Y_{(0)}, Y_{(1)})$. Orchard and Woodbury (1972) give a slightly different form of this expression.

Example 8.5 A Multinomial Example (Example 8.2 Continued). For the multinomial Example 8.2, the complete-data loglikelihood is

$$\ell(\theta | Y) = y_1 \ln(1 - \theta) + (y_2 + y_3) \ln \theta$$

ignoring terms not involving θ . Differentiating $\ell(\theta | Y)$ once and twice with respect to θ yields

$$D_\ell(\theta | Y) = -y_1/(1 - \theta) + (y_2 + y_3)/\theta;$$

$$I(\theta | Y) = y_1/(1 - \theta)^2 + (y_2 + y_3)/\theta^2.$$

Hence

$$E\{I(\theta | Y) | Y_{(0)}, \theta\} = y_1/(1 - \theta)^2 + (y_2 + \hat{y}_3)/\theta^2,$$

$$E\{D_\ell^2(\theta | Y) | Y_{(0)}, \theta\} = \text{Var}\{D_\ell(\theta | Y) | Y_{(0)}, \theta\} = V/\theta^2,$$

where $\hat{y}_3 = E(y_3 | Y_{(0)}, \theta) = (y_3 + y_4)(0.25\theta)(0.25\theta + 0.5)^{-1}$, and $V = \text{Var}(y_3 | Y_{(0)}, \theta) = (y_3 + y_4)(0.5)(0.25\theta)(0.25\theta + 0.5)^{-2}$. Substituting $y_1 = 38$, $y_2 = 34$, $y_3 + y_4 = 125$ and $\hat{\theta} = 0.6268$ in these expressions yields

$$\begin{aligned} E\{I(\theta | Y) | Y_{(0)}, \theta\}|_{\theta=\hat{\theta}} &= 435.3, \\ E\{D_\ell^2(\theta | Y) | Y_{(0)}, \theta\}|_{\theta=\hat{\theta}} &= 57.8. \end{aligned}$$

Hence $I(\hat{\theta} | Y_{(0)}) = 435.8 - 57.8 = 377.5$, as can be verified by direct computation. Note that the ratio of missing information to complete information is $57.8/435.3 = 0.1328$, which governs the speed of convergence of EM near $\hat{\theta}$ as confirmed in the last column of Table 8.1.

The decomposition (8.26) of the observed information is particularly simple when the complete data come from the exponential family (8.19). The complete information is $\text{Var}(s(Y) | \theta)$, and the missing information is $\text{Var}(s(Y) | Y_{(0)}, \theta)$. Thus the observed information is

$$I(\theta | Y_{(0)}) = \text{Var}(s(Y) | \theta) - \text{Var}(s(Y) | Y_{(0)}, \theta), \quad (8.32)$$

the difference between the unconditional and conditional variance of the complete-data sufficient statistic. The ratio of the conditional to the unconditional variance determines the rate of convergence in this case.

8.5 Extensions of EM

8.5.1 The ECM Algorithm

There are a variety of important applications where the M step does not have a simple computational form, even when the complete data are from the exponential family (8.19). In such cases, one way to avoid an iterative M step with each EM iteration is to increase the Q function rather than maximize it at each M step, resulting in a GEM algorithm. GEM algorithms increase the likelihood at each iteration, but appropriate convergence is not guaranteed without further specification of the process of increasing the Q function. The ECM algorithm (Meng and Rubin 1993) is a subclass of GEM that is more broadly applicable than EM, but shares its desirable convergence properties.

The ECM algorithm replaces each M step of EM by a sequence of S conditional maximization steps, that is CM steps, each of which maximizes the Q function over θ but with some vector function of θ , say $g_s(\theta)$, fixed at its previous value, for $s = 1, \dots, S$. The general mathematical expressions involve detailed notation, but it is easy to convey the basic idea. Suppose, as in the following example, that the parameter θ is partitioned into subvectors $\theta = (\theta_1, \dots, \theta_S)$. In

many applications, it is useful to take the s th of the CM steps to be maximization with respect to θ_s , with all other parameters held fixed, whence $g_s(\theta)$ is the vector consisting of all the subvectors except θ_s . In this case, the sequence of CM steps is equivalent to a cycle of the complete-data iterated conditional modes algorithm (Besag 1986), which, if the modes are obtained by finding the roots of score functions, can also be viewed as a Gauss–Seidel iteration in an appropriate order (see, for example, Thisted 1988, chapter 4). Alternatively, it may be useful in other applications to take the s th of the CM steps to be simultaneous maximization over all of the subvectors except for θ_s , which is fixed, implying $g_s(\theta) = \theta_s$. Other choices for the functions g_s , perhaps corresponding to different partitions of θ at each CM step, can also be useful, as illustrated by our second example.

Because each CM step increases Q , it is easy to see that ECM is a GEM algorithm and therefore, like EM, monotonically increases the likelihood of θ . Furthermore, when the set of functions g is “space-filling” in the sense of allowing unconstrained maximization over θ in its parameter space, ECM converges to a stationary point under essentially the same conditions that guarantee the convergence of EM. Meng and Rubin (1993) establishes this precisely, but to see this intuitively, suppose that ECM has converged to θ^* in the interior of the parameter space, and that the required derivatives of Q are all well defined; the stationarity of each ECM step implies that the corresponding directional derivatives of Q at θ^* are zero, which, under the space-filling condition on $\{g_s, s = 1, \dots, S\}$, implies that the vector derivative of Q with respect to θ is zero at θ^* , just as with the M step of EM. Thus, as with EM theory, if ECM converges to θ^* , θ^* must be a stationary point of the observed likelihood.

Example 8.6 illustrates ECM in a simple but rather general model in which partitioning the parameter into a location parameter, θ_1 , and a scale parameter, θ_2 , leads to replacement of an iterative M step by two straightforward CM steps, each involving closed-form maximization over one of the parameters while holding the other fixed.

Example 8.6 *A Multivariate Normal Regression Model with Incomplete Data.* Suppose we have n independent observations from the following K -variate normal model

$$y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma), \quad i = 1, \dots, n, \quad (8.33)$$

where X_i is a known $(K \times p)$ design matrix for the i th observation, β is a $(p \times 1)$ vector of unknown regression coefficients, and Σ is a $(K \times K)$ unknown variance–covariance matrix. By specifying particular mean structures and covariance structures, model (8.33) includes important complete-data models such as “seemingly unrelated regressions” (Zellner 1962) and “general repeated measures” (Jennrich and Schluchter 1986), as special cases. It is known,

however, that ML estimation of $\theta = (\beta, \Sigma)$ is generally not in closed form except in special cases, as when $\Sigma = \sigma^2 I_K$, e.g., Szatrowski (1978). This result implies that, generally, the M step of EM is iterative if it is employed used to fit model (8.33) with missing values in the vectors of outcomes $\{y_i\}$.

Consider ML estimation for complete data from the model (8.33) when Σ is unstructured. Joint maximization of β and Σ is not possible in closed form, but if Σ were known, say $\Sigma = \Sigma^{(t)}$, then the conditional ML estimate of β would be the weighted least-squares estimate:

$$\beta^{(t+1)} = \left\{ \sum_{i=1}^n X_i^T (\Sigma^{(t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^n X_i^T (\Sigma^{(t)})^{-1} Y_i \right\}. \quad (8.34)$$

Given $\beta = \beta^{(t+1)}$, the conditional ML estimate of Σ can be obtained directly from the cross products of the residuals:

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \beta^{(t+1)}) (Y_i - X_i \beta^{(t+1)})^T. \quad (8.35)$$

The complete-data log-likelihood function is increased by each conditional maximization, (8.34) and (8.35):

$$\text{CM1 : } \ell(\beta^{(t+1)}, \Sigma^{(t)} | Y) \geq \ell(\beta^{(t)}, \Sigma^{(t)} | Y),$$

$$\text{CM2 : } \ell(\beta^{(t+1)}, \Sigma^{(t+1)} | Y) \geq \ell(\beta^{(t+1)}, \Sigma^{(t)} | Y).$$

With missing data, write, as before $Y = (Y_{(0)}, Y_{(1)})$, where $Y_{(0)}$ is the observed data and $Y_{(1)}$ is the missing data. One iteration of the ECM algorithm consists of one E step and two noniterative CM steps, say CM1 and CM2. More specifically, at iteration $(t + 1)$:

- (a) *E*: The same as the E step of EM, that is, find the conditional expectation of the complete-data sufficient statistics $E(Y_i | Y_{(0)}, \theta^{(t)})$ and $E(Y_i Y_i^T | Y_{(0)}, \theta^{(t)})$, where $\theta^{(t)} = (\beta^{(t)}, \Sigma^{(t)})$. Details are deferred until Section 11.2.1.
- (b) *CM1*: Calculate $\beta^{(t+1)}$ given $\Sigma^{(t)}$ using (8.34), with Y_i replaced by $E(Y_i | Y_{(0)}, \theta^{(t)})$.
- (c) *CM2*: Calculate $\Sigma^{(t+1)}$ given $\beta^{(t+1)}$ using (8.35), with Y_i and $Y_i Y_i^T$ on the right side replaced with $E(Y_i | Y_{(0)}, \theta^{(t)})$ and $E(Y_i Y_i^T | Y_{(0)}, \theta^{(t)})$, respectively.

The resulting ECM algorithm can be viewed as a generalization of iteratively reweighted least squares, e.g., Rubin (1983a), for data with missing values.

The next example concerns log-linear models for contingency tables with missing data; readers not familiar with these models may omit this example

or first review the material in Chapter 13. The example illustrates two additional features of ECM: first, that more than $S > 2$ CM steps may be useful, and second, that the g_s -functions in the constrained maximizations do not have to correspond to a simple partition of the parameter θ . To allow for the fact that estimates of θ can change at each CM step, the estimate of θ for CM step s within iteration t is denoted $\theta^{(t+s/S)}$ for $s = 1, \dots, S$.

Example 8.7 *A Log-Linear Model for Contingency Tables with Incomplete Data.* It is well known that certain log-linear models do not have closed-form ML estimates even with complete data, for example, the well-known iterative algorithm for fitting these kinds of models is iterative proportional fitting (IPF), e.g., Bishop et al. (1975, chapter 3). With incomplete data, ML estimates can be obtained using the ECM algorithm, with the CM steps corresponding to one iteration of IPF applied to the filled-in data from the E step.

In particular, for the no three-way association model for a $2 \times 2 \times 2$ table, let y_{ijk} and θ_{ijk} be respectively the count and probability in cell ijk ($i, j, k = 1, 2$), where the parameter space Ω_θ is the subspace of $\{\theta_{ijk} : i, j, k = 1, 2\}$ such that the three-way association is zero. Let $\theta_{ij(k)} = \theta_{ijk}/\sum_k \theta_{ijk}$ denote the conditional probability of being in cell k of the third factor given that the observation is in cell (i, j) of the two-way table formed by the first two factors, and define $\theta_{i(j)k}$ and $\theta_{(i)jk}$ analogously. Initialize the parameter estimates at the constant table, $\theta_{ijk}^{(0)} = 1/8$ for all i, j, k . At iteration t , let $\{\theta_{ijk}^{(t)}\}$ be current estimated cell probabilities. The complete-data sufficient statistics (as discussed in Section 13.4) are the three sets of two-way margins $\{y_{ij+}\}$, $\{y_{i+k}\}$ and $\{y_{+jk}\}$; let $\{y_{ij+}^{(t)}\}$, $\{y_{i+k}^{(t)}\}$ and $\{y_{+jk}^{(t)}\}$ be current estimates of these margins at iteration t . That is, $y_{ij+}^{(t)} = E(y_{ij+} | y_{\text{obs}}, \theta^{(t)})$, with analogous replacements for y_{i+k} and y_{+jk} . In iteration $(t+1)$, the parameters are updated by applying IPF to the two-way margins, which involves the following three sets of conditional maximizations:

$$\text{CM1 : } \theta_{ijk}^{(t+1/3)} = \theta_{ij(k)}^{(t)} \left(y_{ij+}^{(t)} / n \right), \quad (8.36)$$

$$\text{CM2 : } \theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)} \left(y_{i+k}^{(t)} / n \right), \quad (8.37)$$

$$\text{CM3 : } \theta_{ijk}^{(t+3/3)} = \theta_{(i)jk}^{(t+2/3)} \left(y_{+jk}^{(t)} / n \right), \quad (8.38)$$

where n is the total count. It is easy to see that (8.36) corresponds to maximizing the log-likelihood $\ell(\theta | Y^{(t)})$ subject to the constraints $\theta_{ij(k)} = \theta_{ij(k)}^{(t)}$ for all i, j, k . Similarly, expressions (8.37) and (8.38) correspond to maximizing the log-likelihood $\ell(\theta | Y^{(t)})$ subject to $\theta_{i(j)k} = \theta_{i(j)k}^{(t+1/3)}$ and $\theta_{(i)jk} = \theta_{(i)jk}^{(t+2/3)}$, respectively. IPF is easy because (i) the constraint of “no three-way association” only

imposes restrictions on the conditional probabilities $\theta_{ij(k)}$, $\theta_{i(j)k}$, and $\theta_{(i)jk}$, and thus, once these conditional probabilities are given, the conditional ML estimates for the two-way marginal probabilities θ_{ij+} , θ_{i+k} , and θ_{+jk} are simply the sample proportions, and (ii) if $\theta^{(0)} \in \Omega_\theta$, then all $\theta^{(t)} \in \Omega_\theta$, so starting from a table of constant probabilities yields the appropriate ML estimates. The details of the E step are deferred until Chapter 13.

The next example is an extension of the EM algorithm in Example 8.4:

Example 8.8 Univariate t with Unknown Degrees of Freedom (Example 8.4 Continued). In Example 8.4, we described an EM algorithm for a random sample from the t distribution with known df, v . It is also possible to estimate simultaneously the location and scale parameters and v , a form of adaptive robust estimation in that the data are used to estimate the parameter v that controls the degree of down-weighting of outliers. The ECM algorithm can be used to provide ML estimates in this case. As in Example 8.4, the complete data are $Y = (X, W)$ and the E step computes the expected values of the complete-data sufficient statistics with estimated weight given by (8.24). The M step is complicated by the estimation of v . ECM exploits the simplicity of the M step when v is known by replacing the M step by the following two CM steps:

CM1: For current parameters $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)}, v^{(t)})$, maximize the Q -function with respect to (μ, σ) with $v = v^{(t)}$. This step involves Eq. (8.25) with v set to its current estimate, $v^{(t)}$, and gives $(\mu, \sigma) = (\mu^{(t+1)}, \sigma^{(t+1)})$.

CM2: Maximize the Q -function (8.15) with respect to v , with $(\mu, \sigma) = (\mu^{(t+1)}, \sigma^{(t+1)})$. Specifically, the complete-data loglikelihood given $Y = (X, W)$ is

$$\begin{aligned} \ell(\mu, \sigma^2, v \mid Y) &= -0.5n \log \sigma^2 - 0.5 \sum_{i=1}^n w_i (x_i - \mu)^2 / \sigma^2 \\ &\quad + 0.5nv \log(v/2) - n \log \Gamma(v/2) + (v/2 - 1) \sum_{i=1}^n \log w_i - 0.5v \sum_{i=1}^n w_i \end{aligned}$$

so the Q -function is

$$\begin{aligned} Q(\mu, \sigma^2, v \mid \mu^{(t)}, \sigma^{(t)2}, v^{(t)}) &= -0.5n \log \sigma^2 - 0.5 \sum_{i=1}^n w_i^{(t)} (x_i - \mu)^2 / \sigma^2 \\ &\quad + 0.5nv \log(v/2) - n \log \Gamma(v/2) + (v/2 - 1) \sum_{i=1}^n z_i^{(t)} - 0.5v \sum_{i=1}^n w_i^{(t)}, \end{aligned} \tag{8.39}$$

where $w_i^{(t)}$ is given by (8.24) and $z_i^{(t)} = E(\log w_i \mid x_i, \mu^{(t)}, \sigma^{(t)2}, v^{(t)})$ involves digamma functions. Note that (8.39) is a complex function of v , but the maximization is confined to a single scalar parameter, and $v^{(t+1)}$ can be found

by an iterative one-dimensional search. A drawback is that convergence is considerably slowed by estimation of v ; the ECME algorithm described next provides a way to speed convergence without increasing the complexity of the algorithm.

8.5.2 The ECME and AECM Algorithms

The Expectation/Conditional Maximization Either (ECME) algorithm (Liu and Rubin 1994) replaces some of the CM steps of ECM, which maximize the constrained expected complete-data loglikelihood function, with steps that maximize the correspondingly constrained actual likelihood function. This algorithm shares with both EM and ECM their stable monotone convergence and basic simplicity of implementation relative to competing faster converging methods. Moreover, ECME can have a substantially faster convergence rate than either EM or ECM, measured using either the number of iterations or actual computer time. There are two reasons for this improvement. First, in some of ECME's maximization steps, the actual likelihood is being conditionally maximized, rather than an approximation of it, as with EM and ECM. Second, ECME allows faster converging numerical methods to be used on only those constrained maximizations where they are most efficacious. Also, the faster rate of convergence allows easier assessment of convergence.

As with EM and ECM, the derivative of the $\theta^{(t)} \rightarrow \theta^{(t+1)}$ mapping for ECME at θ^* governs the convergence rate of ECME, and can be obtained in terms of the missing-data, observed-data, and complete-data information matrices. The mathematical expressions are tedious, but confirm that typically ECME will converge faster than ECM or EM. The intuition is direct because CM steps that maximize L rather than Q maximize the correct functions rather than a current approximation to it. Of course, a special case of ECME is that all steps maximize L with no E steps, which has quadratic convergences. More precisely, the method of Jamshidian and Jennrich (1993) can be viewed technically as a special case of ECME where each of the CM steps maximizes the actual likelihood and the constraint functions correspond to different conjugate linear combinations of the parameters across iterations.

Example 8.9 Univariate t with Unknown Degrees of Freedom (Example 8.8 Continued). In Example 8.8, we described an ECM algorithm for a random sample from the t distribution with unknown df v . An ECME algorithm is obtained by retaining the E and CM1 steps of the algorithm in Example 8.8, but replacing the CM2 step, maximization of (8.39) with respect to v , by maximization of the observed loglikelihood, the sum of the logarithm of (8.22) over the observations i , with respect to v , fixing $(\mu, \sigma) = (\mu^{(t+1)}, \sigma^{(t+1)})$. As with ECM this step involves a one-dimensional search to find $v^{(t+1)}$, but the algorithm converges considerably faster than ECM.

The “alternating expectation conditional maximization” (AECM) algorithm (Meng and Van Dyk 1997) builds on the ECME idea by maximizing functions other than Q or L in particular CM steps, corresponding to varying definitions of what constitutes missing data; maximizing L is the special case of no missing data. An iteration of AECM consists of cycles, each consisting of an E step with a particular definition of complete and missing data, followed by CM steps that correspond to that definition; a set of such cycles that are space-filling maximizations in the sense of ECM parameterizations define one full iteration of AECM. As with ECME, this can result in enhanced computational efficiency.

8.5.3 The PX-EM Algorithm

Parameter-expanded expectation–maximization (PX-EM) (Liu et al. 1998) speeds EM by embedding the model of interest within a larger model with an additional parameter α , such that the original model is obtained by setting α to a particular value, α_0 . If the parameter in the original model is θ , then the parameters of the expanded model are $\phi = (\theta^*, \alpha)$ where θ^* is the same dimension as θ , $\theta = R(\theta^*, \alpha)$ for some known transformation R , with the restriction that $\theta^* = \theta$ when $\alpha = \alpha_0$. The expanded model is chosen so that (i) there is no information about α in the observed data $Y_{(0)}$, that is

$$f_x(Y_{(0)} | \theta^*, \alpha) = f_x(Y_{(0)} | \theta^*, \alpha') \text{ for all } \alpha, \alpha' \quad (8.40)$$

where f_x denotes the density of the expanded model, and (ii) the parameters ϕ in the expanded model have unique ML estimates, with complete data $Y = (Y_{(0)}, Y_{(1)})$. The PX-EM algorithm is simply EM applied to the expanded model; that is, for the t -th iteration:

PX-E step: Compute $Q(\phi | \phi^{(t)}) = E(\log f_x(Y | \phi) | Y_{(0)}, \phi^{(t)})$

PX-M step: Find $\phi^{(t+1)} = \arg \max_{\phi} Q(\phi | \phi^{(t)})$, and then set $\theta^{(t+1)} = R(\theta^{*(t+1)}, \alpha)$.

The theory of EM applied to the expanded model implies that each step of PX-EM increases $f_x(Y_{(0)} | \theta^*, \alpha)$, which equals $f(Y_{(0)} | \theta)$ when $\alpha = \alpha_0$. Hence, each step of PX-EM increases the relevant likelihood $f(Y_{(0)} | \theta)$, and convergence properties of PX-EM parallel that of standard EM, except that PX-EM converges faster; see the discussion following Example 8.10.

Example 8.10 *PX-EM Algorithm for the Univariate t with Known Degrees of Freedom (Example 8.4 Continued).* In Example 8.4, we applied EM to compute ML estimates for the univariate t model (8.22) with known df v by embedding

the observed data X in a large data set (X, W) from the model (8.23). Suppose we replace this model by the expanded complete-data model:

$$(x_i \mid \mu_*, \sigma_*, \alpha, w_i) \sim_{\text{ind}} N(\mu_*, \sigma_*^2/w_i), \quad (w_i \mid \mu_*, \sigma_*, \alpha) \sim_{\text{ind}} \alpha \chi_v^2/v, \quad (8.41)$$

where $\theta^* = (\mu_*, \sigma_*)$ and α is an additional scale parameter. The expanded model (8.41) reduces to the original model (8.23) when $\alpha = 1$. Because the marginal density (8.22) of the observed data X is unchanged and does not involve α , there is no information about α in $Y_{(0)} = X$, but α has a unique ML estimate from the complete data (X, W) . So both conditions for applying PX-EM are satisfied. The transformation R from (θ^*, α) to θ is $\mu = \mu_*$, $\sigma = \sigma_*/\sqrt{\alpha}$. The PX-E step is similar to the E step (8.24) of EM:

PX-E step: At iteration $(t + 1)$, compute:

$$w_i^{(t)} = E(w_i \mid x_i, \phi^{(t)}) = \alpha^{(t)} \frac{v + 1}{v + d_i^{(t)2}}, \quad (8.42)$$

where $d_i^{(t)} = \sqrt{\alpha^{(t)}}(x_i - \mu_*^{(t)})/\sigma_*^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)}$, as in (8.24).

The PX-M step maximizes the expected complete-data log-likelihood of the expanded model:

PX-M step: Compute $\mu_*^{(t+1)}, \sigma_*^{(t+1)}$ as for the M step (8.25) of EM, that is

$$\mu_*^{(t+1)} = \frac{s_1^{(t)}}{s_0^{(t)}}, \quad \sigma_*^{(t+1)2} = \frac{1}{n} \sum_{i=1}^n w_i^{(t)} (x_i - \mu_*^{(t+1)})^2 = \frac{s_2^{(t)} - s_1^{(t)2}/s_0^{(t)}}{n},$$

and

$$\alpha^{(t+1)} = \frac{1}{n} \sum_{i=1}^n w_i^{(t)}.$$

In the original parameter space, the PX-M step is

$$\mu^{(t+1)} = \frac{s_1^{(t)}}{s_0^{(t)}}; \quad \sigma^{(t+1)2} = \sum_{i=1}^n w_i^{(t)} (x_i - \mu_*^{(t+1)})^2 / \sum_{i=1}^n w_i^{(t)}. \quad (8.43)$$

Thus in practical terms, the modification of EM is simply to use the sum of the weights in the denominator of the estimate of σ^2 rather than the sample size. This modification was previously proposed by Kent et al. (1994) as a modification of EM to speed convergence, but without the more general PX-EM motivation.

To understand why PX-EM converges more rapidly than the original EM algorithm in terms of θ , it is clearer to reparameterize from $\phi = (\theta^*, \alpha)$ to (θ, α) , where $\theta = R(\theta^*, \alpha)$ is the appropriate transformation. Equation (8.40) implies that the observed-data and complete-data information matrices for $f_x(Y | \theta, \alpha)$ in this parameterization are

$$I_{\text{obs}} = \begin{pmatrix} i_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad I_{\text{com}} = \begin{pmatrix} i_{\text{com}} & i_{\theta\alpha}^T \\ i_{\theta\alpha} & i_{\alpha\alpha} \end{pmatrix}.$$

Thus, the fraction of missing information (8.28) is the (θ, θ) submatrix of $I - I_{\text{obs}}I_{\text{com}}^{-1}$, namely

$$U(\theta) = I - i_{\text{obs}}(i_{\text{com}} - i_{\theta\alpha}^T i_{\alpha\alpha}^{-1} i_{\theta\alpha})^{-1},$$

which is smaller than the fraction of missing information in the original model, $I - i_{\text{obs}}i_{\text{com}}^{-1}$, in the sense that the difference in these two matrices is semi-negative definite. Because the fraction of missing information is smaller, the rate of convergence of PX-EM is faster than EM for the original model. In other words, the effect of expanding the model to include α has been to reduce the complete-data information about θ from i_{com} to $i_{\text{com}} - i_{\theta\alpha}^T i_{\alpha\alpha}^{-1} i_{\theta\alpha}$ without changing the observed-data information i_{obs} about θ . This has the effect of reducing the fraction of missing information and hence speeding EM.

The practical gains of PX-EM in speeding EM are modest in this illustrative example, but they become more substantial in generalizations of the model to multivariate incomplete X , as considered in Chapter 14.

8.6 Hybrid Maximization Methods

The slow convergence of EM-type algorithms has motivated a variety of attempts to speed the algorithm by combining it with Newton–Raphson or Scoring-type updates, or Aitken acceleration (Louis 1982; Meilijson 1989; McLachlan and Krishnan 1997, section 4.7), a variant of Newton–Raphson. A simple version of this idea is to combine EM steps and Newton steps in a way that exploits the advantages of both. For example, a Newton step might be attempted, and replaced by one or more EM steps if the likelihood is not increased. A reasonable hybrid approach starts with EM steps, when the parameters are far from ML and Newton steps are more likely to fail, and finishes with Newton steps, because the loglikelihood near the maximum may be closer to quadratic, and final convergence of EM may be hard to determine if the EM steps are small. Of course, these approaches are not useful when EM is used to avoid the programming complexity of the Newton-type methods.

When the M step of EM requires iteration, one option is to replace the M step by one Newton step applied to the Q-function, a method known as gradient EM (Lange 1995a). More generally, Lange (1995a) considers algorithms of the form:

$$\theta^{(t+1)} = \theta^{(t)} + a^{(t)} \delta^{(t)}, \quad (8.44)$$

where

$$\delta^{(t)} = \left(-D^{20}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}} \right)^{-1} D^{10}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}}, \quad (8.45)$$

where $a^{(t)}$ is a constant between 0 and 1 chosen so that

$$Q(\theta^{(t+1)}; \theta^{(t)}) \geq Q(\theta^{(t)}; \theta^{(t)}), \quad (8.46)$$

that is, so that (8.44) defines a GEM algorithm. Gradient EM is a special case of (8.44) and (8.45) with $a^{(t)} = 1$. In many cases $-D^{20}Q(\theta; \theta)$ is positive definite, in which case (8.46) can always be achieved by choosing $a^{(t)}$ to be sufficiently small, for example by successive step-halving.

In Lange's (1995b) quasi-Newton acceleration method, the update (8.44) is applied with (8.45) replaced with

$$\delta^{(t)} = \left(-D^{20}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}} + B^{(t)} \right)^{-1} D^{10}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}}, \quad (8.47)$$

where the adjustment $B^{(t)}$ is chosen to bring $-D^{20}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}} + B^{(t)}$ closer to the Hessian matrix $-D^{20}\ell(\theta | Y_{\text{obs}})_{\theta=\theta^{(t)}}$ of Newton–Raphson applied directly to the observed-data likelihood.

Lange proposes a choice of B that only involves the first derivatives of the Q-function: $B^{(0)} = 0$ and

$$B^{(t)} = B^{(t-1)} - (\nu^{(t)} \nu^{(t) T}) / (\nu^{(t) T} (\theta^{(t)} - \theta^{(t-1)})),$$

$$\nu^{(t)} = h^{(t)} + B^{(t-1)}(\theta^{(t)} - \theta^{(t-1)}),$$

$$h^{(t)} = D^{10}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t-1)}} - D^{10}Q(\theta; \theta^{(t-1)})_{\theta=\theta^{(t-1)}}.$$

If $-D^{20}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}} + B^{(t)}$ in (8.47) is not positive definite, Lange proposes replacing it by $-D^{20}Q(\theta; \theta^{(t)})_{\theta=\theta^{(t)}} + B^{(t)} / 2^m$, where m is the smallest positive integer that yields a positive-definite matrix. Similarly, if the choice $a^{(t)} = 1$ in (8.44) does not yield an increase in loglikelihood, the step $a^{(t)}$ can be halved repeatedly until an increase is obtained. Lange (1995b) notes that this algorithm resembles EM in its early stages and Newton–Raphson at its later stages, and at intermediate stages makes a graceful transition between these two extremes.

The methods discussed so far in this section all require computation and inversion of a $(q \times q)$ matrix at each iteration, which is potentially time-consuming when the number of parameters is large. Jamshidian and Jennrich (1993) propose an acceleration of EM that avoids this inversion, based on generalized conjugate-gradient ideas, which they call the accelerated EM algorithm.

This algorithm appends to EM a line search in the direction of the change in EM iterates (see Jamshidian and Jennrich (1993) for details).

Problems

- 8.1** Show that for a scalar parameter, the Newton–Raphson algorithm converges in one step if the loglikelihood is quadratic.
- 8.2** Describe in words the purpose of the E and M steps of the EM algorithm.
- 8.3** Prove that the loglikelihood in Example 8.3 is linear in the statistics in Eq. (8.12).
- 8.4** Show how Corollaries 8.1 and 8.2 follow from Theorem 8.1.
- 8.5** Review results concerning the convergence of EM.
- 8.6** Show that (8.20) and (8.21) are the E and M steps for the regular exponential family (8.19).
- 8.7** Suppose $Y = (y_1, \dots, y_n)^T$ are independent gamma random variables with density

$$f(y_i | k, \theta_i) = x^{k-1} \exp(-y_i/\theta_i) / (\theta_i^k \Gamma(k)),$$

with unknown shape k , scale θ_i , and mean $k\theta_i = g(\sum_j \beta_j x_{ij})$, where g is a known function, $\beta = (\beta_1, \dots, \beta_J)$ are unknown regression coefficients, and x_{i1}, \dots, x_{ij} are the values of covariates X_1, \dots, X_J for unit i . For what choice of g does Y belong to the regular $J+1$ parameter exponential family, and what are the natural parameters and complete-data sufficient statistics?

- 8.8** Suppose values y_i in Problem 8.7 are missing if and only if $y_i > c$, for some known censoring point c . Explore the E step of the EM algorithm for estimating (a) β_1, \dots, β_J when k is known; (b) β_1, \dots, β_J and k , when k is unknown.
- 8.9** By hand calculation, carry out the multivariate normal EM algorithm for the data set in Table 7.1, with initial estimates based on the complete observations. Hence verify that, for this pattern of data and choice of starting values, the algorithm converges after one iteration (i.e., subsequent iterations lead to the same answer as the first iteration). Why does Eq. (8.29) not apply in this case? (*Hint:* Consider Corollary 8.2 of Theorem 8.1 with $\theta^{(t)} = \theta^*$.)

- 8.10** Write the loglikelihood of θ for the observed data in Example 8.2. Show directly by differentiating this function that $I(\theta | Y_{(0)}) = 435.3$, as found in Example 8.5.
- 8.11** Verify the E and M steps in Example 8.4.
- 8.12** Write the asymptotic sampling variance of the ML estimate of θ in Example 8.2, and compare it with the sampling variance of the ML estimate when the first and third counts (namely, 38 and 125) are combined, yielding counts (163, 34) from a binomial distribution with probabilities $(1 - \theta/4, \theta/4)$.
- 8.13** For the censored exponential sample in the second part of Example 6.22, suppose y_1, \dots, y_r are observed and y_{r+1}, \dots, y_n are censored at c . Show that the complete-data sufficient statistic for this problem is $s(Y) = \sum_{i=1}^n y_i$, and the natural parameter is $\phi = 1/\theta$, the reciprocal of the mean. Find the observed information for ϕ by computing the unconditional and conditional variance of $s(Y)$ and subtracting, as in (8.32). Hence, find the proportion of missing information from the censoring, and the asymptotic sampling variances of $\hat{\phi} - \phi$ and $\hat{\theta} - \theta$.
- 8.14** Write the complete-data loglikelihood in Example 8.6, and verify the two CM steps, Eqs. (8.34) and (8.35), in that example.
- 8.15** Prove the PX-E and PX-M steps in Example 8.10.
- 8.16** Suppose that (a) X is Bernoulli with $\Pr(X = 1) = 1 - \Pr(X = 0) = \pi$, and (b) Y given $X = j$ is normal with mean μ_j , variance σ^2 , a simple form of the discriminant analysis model. Consider now the monotone missing-data pattern with Y completely observed but $n - r$ values of X missing, and an ignorable missingness mechanism. In Problem 7.17, we found that the factored likelihood method of Chapter 7 does not provide closed-form expressions for ML estimates. Describe the E and M steps of the EM algorithm for this problem, and provide a flow chart for programming this algorithm.

9

Large-Sample Inference Based on Maximum Likelihood Estimates

9.1 Standard Errors Based on The Information Matrix

We noted in Chapter 6 that large-sample maximum likelihood (ML) inferences can be based on Approximation 6.1, namely that

$$(\theta - \hat{\theta}) \sim N(0, C), \quad (9.1)$$

where C is an estimate of the $d \times d$ covariance matrix of $(\theta - \hat{\theta})$, for example,

$$C = I^{-1}(\hat{\theta} | Y_{(0)}), \quad (9.2)$$

the inverse of the observed information evaluated at $\theta = \hat{\theta}$, or

$$C = J^{-1}(\hat{\theta}), \quad (9.3)$$

the inverse of the expected information evaluated at $\theta = \hat{\theta}$, or

$$\hat{C}^* = I^{-1}(\hat{\theta}) \hat{K}(\hat{\theta}) I^{-1}(\hat{\theta}), \quad \text{where } \hat{K}(\hat{\theta}) = \left. \frac{\partial \ell(\theta | Y_{(0)})}{\partial \theta} \frac{\partial \ell(\theta | Y_{(0)})}{\partial \theta}^T \right|_{\theta=\hat{\theta}}, \quad (9.4)$$

the sandwich estimator. The estimate (9.2) is computed as part of the Newton–Raphson algorithm for ML estimation, and (9.3) is computed as part of the scoring algorithm. When the expectation–maximization (EM) algorithm or one of the variants described in Chapter 8 is used for ML estimation, additional steps are needed to compute standard errors of the estimates.

The estimate of the observed information matrix $I(\hat{\theta} | Y_{(0)})$ in (9.2) can be found directly by differentiating the loglikelihood $\ell(\theta | Y_{(0)})$ twice with respect

to θ . Alternatively, it can be computed as the difference of the complete information and missing information using

$$I(\theta \mid Y_{(0)}) = -D^{20}Q(\theta \mid \theta) + D^{20}H(\theta \mid \theta), \quad (9.5)$$

or one of the similar expressions in Chapter 8. The next section considers methods for computing standard errors that do not require computation and inversion of an information matrix.

9.2 Standard Errors via Other Methods

9.2.1 The Supplemented EM Algorithm

Supplemented expectation–maximization (SEM) (Meng and Rubin 1991) is a way to calculate the large-sample covariance matrix associated with $\theta - \hat{\theta}$ using only (i) code for the E and M steps of EM, (ii) code for the large-sample complete-data variance–covariance matrix, V_c , and (iii) standard matrix operations. In particular, no further mathematical analysis of the specific problem is needed beyond that needed for the complete-data large-sample inference (namely the M step and V_c), and that needed for the E step. SEM tends to be more computationally stable than numerically differentiating $\ell(\theta \mid Y_{(0)})$, because the numerical approximations are applied only to the missing information, using analytical expressions for the complete-data information matrix.

Recall from Chapter 8 that

$$DM = i_{\text{mis}} i_{\text{com}}^{-1} = I - i_{\text{obs}} i_{\text{com}}^{-1}, \quad (9.6)$$

where DM is the derivative of the EM mapping, $i_{\text{com}} = -D^{20}Q(\theta \mid \theta)|_{\theta=\theta^*}$ is the complete information, $i_{\text{obs}} = I(\theta \mid Y_{(0)})|_{\theta=\theta^*}$ is the observed information, and $i_{\text{mis}} = -D^{20}H(\theta \mid \theta)|_{\theta=\theta^*}$ is the missing information at the converged value of θ . Equation (9.6) implies that $i_{\text{obs}}^{-1} = i_{\text{com}}^{-1}(I - DM)^{-1}$, that is

$$V_{\text{obs}} = V_{\text{com}}(I - DM)^{-1}, \quad (9.7)$$

where $V_{\text{obs}} = i_{\text{obs}}^{-1}$, $V_{\text{com}} = i_{\text{com}}^{-1}$ are variance covariance matrices for the observed data and the complete data, respectively. Hence

$$V_{\text{obs}} = V_{\text{com}}(I - DM + DM)(I - DM)^{-1} = V_{\text{com}} + \Delta V, \quad (9.8)$$

where

$$\Delta V = V_{\text{com}}DM(I - DM)^{-1} \quad (9.9)$$

is the increase in variance due to missing data. The key idea of SEM is that even though M does not have an explicit mathematical form, its derivative, DM , can

be estimated from the output of “forced EM” steps that effectively numerically differentiate M.

Specifically, first obtain the ML estimate $\hat{\theta}$ of θ and then run a sequence of SEM iterations, where iteration $(t + 1)$ is defined as follows:

INPUT: $\hat{\theta}$ and $\theta^{(t)}$.

Step 1: Run the usual E and M steps to obtain $\theta^{(t+1)}$.

Fix $i = 1$. Calculate

$$\theta^{(t)}(i) = (\hat{\theta}_1, \dots, \hat{\theta}_{i-1}, \theta_i^{(t)}, \hat{\theta}_{i+1}, \dots, \hat{\theta}_d),$$

which is $\hat{\theta}$ except in the i th component, which equals $\theta_i^{(t)}$.

Step 3: Treating $\theta^{(t)}(i)$ as the current estimate of θ , run one iteration of EM to obtain $\tilde{\theta}^{(t+1)}(i)$.

Step 4: Obtain the ratio

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \hat{\theta}_j}{\theta_i^{(t)} - \hat{\theta}_i}, \quad \text{for } j = 1, \dots, d.$$

Step 5: Repeat Steps 2–4 for $i = 2, \dots, d$.

OUTPUT: $\theta^{(t+1)}$ and $\{r_{ij}^{(t)} : i, j = 1, \dots, d\}$.

DM is the limiting matrix $\{r_{ij}\}$ as $t \rightarrow \infty$; the element r_{ij} is obtained when the sequence $r_{ij}^{(t^*)}, r_{ij}^{(t^*+1)}, \dots$ is stable for some t^* . This process can result in using different values of t^* for different r_{ij} elements. When all elements in the i th row of DM have been obtained, there is no need to repeat the above Steps 2–4 for that i in subsequent iterations.

Example Step 2:9.1 Standard Errors for Multinomial Data (Example 8.5 Continued). Consider the multinomial data of Example 8.2, with observed counts $Y_{(0)} = (38, 34, 125)$ from a multinomial distribution with cell probabilities $(1/2 - 1/2\theta, 1/4\theta, 1/2 + 1/4\theta)$. In Example 8.2, we applied EM, where the complete data are counts $Y_{\text{com}} = (Y_1, Y_2, Y_3, Y_4)^T$ from a multinomial distribution with parameters $(1/2 - 1/2\theta, 1/4\theta, 1/4\theta, 1/2)$ and $Y_{(0)} = (y_1, y_2, y_3 + y_4)$. EM yielded $\hat{\theta} = 0.6268$ (see Table 8.1). In this case θ is a scalar, and SEM has a particularly simple form, because its standard error can be obtained directly from the EM computations. The complete-data estimator of θ is $(y_2 + y_3)/(y_1 + y_2 + y_3)$ and its complete-data variance is

$$V_{\text{com}} = \hat{\theta}(1 - \hat{\theta})/(y_1 + y_2 + y_3) = 0.6268(1 - 0.6268)/101.83 = 0.002297,$$

where the denominator is the expected value of $y_1 + y_2 + y_3$, given $\hat{\theta}$. The rate of convergence of EM is DM = 0.1328, as seen in the last column of Table 8.1. Hence, from Eq. (9.7), the large-sample variance of $\hat{\theta}$ is

$$V_{\text{obs}} = V_{\text{com}} / (1 - \text{DM}) = 0.002\,297 / (1 - 0.1328) = 0.002\,65.$$

The observed information by analytical calculation is $I_{\text{obs}} = 377.5$, as shown in Example 8.5. Inverting this quantity $V_{\text{obs}} = 1/377.5 = 0.002\,65$, which agrees with the SEM computation. When EM and SEM are applied to $\text{logit}(\theta)$, which should satisfy the assumptions of asymptotic normality better, we find $\text{logit}(\hat{\theta}) = 0.5186$, with associated large-sample variance 0.4841.

When there is no missing information on a particular set of components of θ , EM will converge in one step for those components from any starting value. Hence, the above method needs modification. Suppose the first d_1 components of θ have no missing information. Meng and Rubin (1991) show that the DM matrix has the form

$$\text{DM} = \begin{matrix} d_1 & d_2 \\ \begin{matrix} d_1 \\ d_2 \end{matrix} & \begin{pmatrix} 0 & A \\ 0 & \text{DM}^* \end{pmatrix} \end{matrix}, \quad d_1 + d_2 = d, \quad (9.10)$$

and DM^* can be computed by running Steps 2–4 for $i = d_1 + 1, \dots, d$. Writing

$$V_{\text{com}} = I_{\text{com}}^{-1} = \begin{matrix} d_1 & d_2 \\ \begin{matrix} d_1 \\ d_2 \end{matrix} & \begin{pmatrix} G_1 & G_2 \\ G_2^T & G_3 \end{pmatrix} \end{matrix}, \quad (9.11)$$

the large-sample covariance matrix of $\hat{\theta}$ can be computed via the following generalizations of Eqs. (9.8) and (9.9):

$$V_{\text{obs}} = \begin{pmatrix} G_1 & G_2 \\ G_2^T & G_3 + \Delta V^* \end{pmatrix}, \quad (9.12)$$

where

$$\Delta V^* = (G_3 - G_2^T G_1^{-1} G_2) \text{DM}^* (I - \text{DM}^*)^{-1}. \quad (9.13)$$

Example 9.2 *Standard Errors for a Bivariate Normal Sample with Monotone Missing Data (Example 7.6 Continued).* We illustrate SEM using the data given in Table 7.1, which are assumed, as in Examples 7.2 and 7.3, to follow a bivariate

Table 9.1 Example 9.2, ML estimates for data in Table 7.1, and asymptotic standard errors (SE)

Parameter	μ_2	$\ln \sigma_{22}$	Z_ρ
ML estimate ($\theta_2^{(65)}$)	49.33	4.74	-1.45
SE from Table 7.2	2.73	0.37	0.274
SE from SEM	2.73	0.37	0.274

normal distribution with parameters $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \rho)$, where ρ is the correlation coefficient. As is well known, a normalizing parameterization in this case is $\theta = (\mu_1, \mu_2, \ln \sigma_{11}, \ln \sigma_{22}, Z_\rho)$, where $Z_\rho = 0.5 \ln((1 + \rho)/(1 - \rho))$ is the Fisher Z transformation of ρ . Because the first variable is fully observed, the ML estimates for μ_1 and $\ln \sigma_{11}$ are simply the sample mean and the log of the sample variance (with divisor n) of the first variable, respectively. Thus EM will converge in one step for these two components from any starting values, with the result that the corresponding components of $M(\theta)$ are constant functions. The implementation of EM for the multivariate normal distribution using the sweep operator is described in Section 11.2.

The first row of Table 9.1 gives the ML estimate for $\theta_2 = (\mu_2, \ln \sigma_{22}, Z_\rho)$ using $\theta_2^* = \theta_2^{(65)}$. (In this case, the closed-form value of θ_2^* can be obtained by factoring the likelihood, as in Section 7.2.1.) The second row gives asymptotic standard errors for $\theta_2 - \theta_2^*$, obtained by direct computation as in Section 7.2.2 (and transformed via the appropriate Jacobian), and the third row gives the corresponding standard errors obtained by SEM.

The SEM results are obtained as follows, using the method described above. First, we obtain DM^* , the submatrix of DM corresponding to $\theta_2 = (\mu_2, \ln \sigma_{22}, Z_\rho)$ in Eq. (9.9). Because the complete-data distribution is from a regular exponential family (the standard bivariate normal), to obtain I_{com}^{-1} we only need to compute the inverse of the complete-data information matrix $I^{-1}(\theta^*)$, which is particularly easy to do for the bivariate normal distribution. We find

$$I_{\text{com}}^{-1} = I^{-1}(\theta^*) = \ln \sigma_{11} \begin{pmatrix} \mu_1 & \mu_2 & \ln \sigma_{11} & \ln \sigma_{22} & Z_\rho \\ \mu_1 & 4.9741 & -5.0387 & 0 & 0 \\ \mu_2 & -5.0387 & 6.3719 & 0 & 0 \\ \ln \sigma_{22} & 0 & 0 & 0.1111 & 0.0890 & -0.0497 \\ Z_\rho & 0 & 0 & 0.0890 & 0.1111 & -0.0497 \\ & 0 & 0 & -0.0497 & -0.0497 & 0.0556 \end{pmatrix}. \quad (9.14)$$

After a rearrangement that makes the first two rows and columns correspond to the parameters of the first component, for which there is no missing information, the right side of (9.14) becomes

$$\begin{matrix} \mu_1 & \ln \sigma_{11} & \mu_2 & \ln \sigma_{22} & Z_\rho \\ \mu_1 & \left(\begin{array}{ccccc} 4.9741 & 0 & -5.0387 & 0 & 0 \\ 0 & 0.1111 & 0 & 0.0890 & -0.0497 \\ -5.0387 & 0 & 6.3719 & 0 & 0 \\ 0 & 0.0890 & 0 & 0.1111 & -0.0497 \\ 0 & -0.0497 & 0 & -0.0497 & 0.0556 \end{array} \right) & = & \begin{pmatrix} G_1 & G_2 \\ G_2^T & G_3 \end{pmatrix}, \\ \ln \sigma_{11} \\ \mu_2 \\ \ln \sigma_{22} \\ Z_\rho \end{matrix} \quad (9.15)$$

where G_3 is the (3×3) lower right submatrix of (9.15). Applying formula (9.13), we obtain

$$\Delta V^* = \ln \sigma_{22} \begin{pmatrix} \mu_2 & \ln \sigma_{22} & Z_\rho \\ \mu_2 & \left(\begin{array}{ccc} 1.0858 & 0.1671 & -0.0933 \\ 0.1671 & 0.0286 & -0.0098 \\ -0.0933 & -0.0098 & 0.0194 \end{array} \right) & \\ \ln \sigma_{22} \\ Z_\rho \end{pmatrix}, \quad (9.16)$$

which is the increase in the variance of $\theta_2 - \theta_2^*$ due to missing information. To obtain the asymptotic variance–covariance matrix for $\theta_2 - \theta_2^*$, we only need to add ΔV^* to G_3 of (9.15). For example, for the standard error of $\mu_2 - \mu_2^*$, we have, from (9.14) and (9.15), $(6.3719 + 1.0858)^{1/2} \approx 2.73$, as given in the third row of Table 9.1.

An attractive feature of SEM is that the final answer, V_{obs} , is typically very stable numerically for the following reasons. When the fractions of missing information are small, ΔV is small relative to V_{com} , and although the calculation of DM (used to calculate ΔV) is subject to substantial numerical inaccuracy because of the rapid convergence of EM, this has little effect on the calculated $V_{\text{obs}} = V_{\text{com}} + \Delta V$. When the fractions of missing information are large, ΔV is an important component of V_{com} , but then the relatively slower convergence of EM ensures relatively accurate numerical differentiation of M.

Another attractive feature of SEM is that it produces internal diagnostics for programming and numerical errors. In particular, ΔV is analytically symmetric but may not be numerically symmetric due to programming errors or insufficient numerical precision in the calculation of $\hat{\theta}$ or in DM. Hence, asymmetry of the estimated covariance matrix from SEM is an indication of a programming error, possibly in the original EM algorithm. Furthermore, even if symmetric,

V_{obs} may not be positive semidefinite, again suggesting either programming or numerical errors, or convergence to a saddle point.

SEM has the advantage over alternative approaches of staying inferentially and computationally closer to EM. Whatever method is used to calculate V_{obs} , however, it is practically important to do so using transformations of θ that tend to normalize the likelihood (e.g., log variance rather than variance with normal models), otherwise, the large-sample standard errors and resulting inferences may be misleading (see Example 7.3); furthermore, SEM will converge more quickly and accurately if such transformations are used.

9.2.2 Bootstrapping the Observed Data

In Examples 5.3 and 5.4, we described the bootstrap as an approach to estimating standard errors from imputed data. This approach can also be applied to estimate standard errors of ML estimates (Little 1988b; Efron 1994). Let $\hat{\theta}$ be the ML estimate of θ based on a sample $S = \{i: i = 1, \dots, n\}$ of independent (possibly incomplete) units. Let $S^{(b)}$ be a sample of n units obtained from the original sample S by simple random sampling with replacement; the bootstrap samples are readily generated by assigning a weight $m_i^{(b)}$ to unit i , where

$$(m_1^{(b)}, m_2^{(b)}, \dots, m_n^{(b)}) \sim \text{MNOM}(n; (n^{-1}, n^{-1}, \dots, n^{-1})),$$

a multinomial distribution with sample size n and n cells with equal probabilities $1/n$. Then $m_i^{(b)}$ represents the number of times that unit i is included in the b th bootstrap sample, with $\sum_{i=1}^n m_i^{(b)} = n$. Let $\hat{\theta}^{(b)}$ be the ML estimate of θ based on data $S^{(b)}$, and let $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$ be the set of estimates obtained by repeating this procedure B times. The bootstrap estimate of θ is then the average of the bootstrap estimates:

$$\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{(b)}, \quad (9.17)$$

and the bootstrap estimate of the sampling variance of $\hat{\theta}$ or $\hat{\theta}_{\text{boot}}$ is

$$\hat{V}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2. \quad (9.18)$$

It can be shown that, under quite general conditions, \hat{V}_{boot} is a consistent estimate of the sampling variance of $\hat{\theta}$ or $\hat{\theta}_{\text{boot}}$ as B tend to infinity. If the

bootstrap distribution is approximately normal, a $100(1 - \alpha)\%$ bootstrap interval for a scalar θ can be computed as

$$I_{\text{norm}}(\theta) = \hat{\theta} \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{boot}}}, \quad (9.19)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the normal distribution. Alternatively, if the bootstrap distribution is nonnormal, a $100(1 - \alpha)\%$ bootstrap interval can be computed as

$$I_{\text{emp}}(\theta) = (\hat{\theta}^{(b,l)}, \hat{\theta}^{(b,u)}), \quad (9.20)$$

where $\hat{\theta}^{(b,l)}$ and $\hat{\theta}^{(b,u)}$ are the empirical $(\alpha/2)$ and $(1 - \alpha/2)$ quantiles of the bootstrap distribution of θ . As discussed in Chapter 5, stable intervals based on (9.19) require bootstrap samples of the order of $B = 200$, whereas intervals based on (9.20) require much larger samples, $B = 2000$ or more (Efron 1994).

This approach assumes large samples. With moderate-sized data sets and extensive missing data, it is possible that ML estimates cannot be computed for particular bootstrap samples because some parameters do not have unique ML estimates. These samples can be omitted in the calculation of (9.19) or (9.20) without affecting the asymptotic consistency of the bootstrap standard errors, but the impact on the validity of the procedure in finite samples is not clear. In our limited experience, it appears that omitting these samples can lead to a potentially severe underestimation of sampling variance; see Example 7.9.

An advantage of the bootstrap is that it can provide a valid asymptotic estimate of the standard errors of ML estimates even if the model is misspecified; asymptotically its properties are parallel to that of the sandwich estimator (9.4) of the covariance matrix, which has the same characteristic. More generally, Efron (1994) notes that the bootstrap provides valid asymptotic standard errors regardless of the validity of assumptions of the model, including the assumption in ignorable ML methods that the missingness mechanism is missing at random (MAR). Although in a sense technically true, inferences based on ignorable ML with bootstrap standard errors are misleading if the ML estimate of θ is inconsistent: they yield p -values with the wrong rejection rates under null hypotheses, and confidence intervals, (9.19) or (9.20), with incorrect coverage, because they have the “right” width centered at the “wrong” value. Model assumptions (including MAR) are needed to ensure consistency of the ML estimate, so these assumptions remain crucial when the bootstrap is used to compute sampling variances.

9.2.3 Other Large-Sample Methods

Other methods for computing an asymptotic covariance matrix have been proposed. Two, like SEM, are based on EM-type computations. Louis (1982)

suggests inverting the observed information computed via Eq. (8.30), which requires the code for computing EM and the complete-data covariance matrix, plus new code for calculating the conditional expectation of the squared complete-data score function. Sometimes, this extra calculation is easy, but often it is not. An application is given in Tu et al. (1993), where the missing data are created by the censoring and truncation of AIDS survival times.

A related method due to Meilijson (1989) avoids the extra analytic calculations of Louis's method, but in its direct form requires the restrictive assumption that the *observed* data are independent and identically distributed. Also, because it replaces the theoretical expectation of the observed squared score with the corresponding average in the sample, it relies on the data actually arising from the fitted model for its numerical propriety.

Another method for calculating large-sample covariance matrices involves a two-part quadratic approximation to the log likelihood. First, find some initial approximation (full rank) to the covariance matrix, for example, possibly based on the complete cases. This is used to create a normal distribution centered at the ML estimate, and then to draw a set of values $\{\theta^{(d)}\}$ of θ , concentrated in the region where interest is focused. For instance, if 95% intervals are of primary interest, values are drawn between 1.5 and 2.5 initial standard errors away from the ML estimate. Then a quadratic response surface is fitted, with dependent variable $\ell(\theta^{(d)} | Y_{(0)})$ as a function of the drawn values $\theta^{(d)}$. Although this method assumes large-sample normality, it should be less sensitive to small-sample abnormalities close to the ML estimate than methods based directly on the information matrix, either calculated analytically or estimated numerically.

A final method for obtaining a large-sample covariance matrix is multiple imputation, introduced in Section 5.4. This technique is addressed in some detail in the next chapter.

9.2.4 Posterior Standard Errors from Bayesian Methods

Another way of computing standard errors without inverting an information matrix is to carry out a Bayesian analysis with a diffuse prior distribution and to use posterior standard deviations to estimate standard errors. Bayes methods for incomplete data are discussed in the following chapter. This approach works because, as noted in Chapter 6, the mode of the posterior distribution from a Bayesian analysis with a uniform prior for the parameters is the ML estimate, and the posterior variance is a consistent estimate of the large-sample variance of the ML estimate because of Approximation 6.1. An advantage of the Bayes approach is that it mimics ML inference in large samples, but also provides Bayesian inference based directly on a posterior distribution, which (in our experience) is often superior to ML in small samples. We discuss this option in more detail in the next chapter.

Problems

- 9.1** In Example 9.1, show how the EM and SEM answers were obtained for $\text{logit}(\hat{\theta})$. Compare the interval estimates for θ using EM/SEM on the raw and logit scales.
- 9.2** Apply SEM to Example 9.2, but without the normalizing transformations on σ_{22} and ρ . Compare the intervals for σ_{22} and ρ based on the results of Example 9.2 and the results in this problem. In theory, which are preferable?
- 9.3** Suppose ECM is used to find the ML estimate of θ in Example 8.6. Further suppose that SEM is applied to the sequence of ECM iterates, assuming they were EM iterates. Are the iterates likely to converge more quickly or more slowly than EM? Would the resulting asymptotic covariance matrix more likely be an over or underestimate, and explain your reasoning. Comment on the possible asymmetry of the calculated matrix? (See Van Dyk et al. 1995, for details.)
- 9.4** Compute standard errors for the data in Table 7.1 using the bootstrap, and compare the results with the standard errors in Table 9.1.
- 9.5** The SEM algorithm can be extended to the SECM algorithm when ECM is used rather than EM. Details are provided in Van Dyk et al. (1995), but it is more complicated than SEM. Describe how the bootstrap can be used to estimate the variance of $(\theta - \hat{\theta})$, where $\hat{\theta}$ is the ML estimate of θ found by ECM.
- 9.6** Suppose PX-EM is used to find the ML estimate of θ in Example 8.10. Further suppose that SEM is applied to the sequence of PX-EM iterates, assuming the algorithm was EM. Would the resulting estimated of standard errors more likely be over or underestimates of the corresponding asymptotic standard errors? Explain your reasoning.
- 9.7** Suppose the model is misspecified, but the ML estimate found by EM is a consistent estimate of the parameter θ . Which method of estimating the large-sample covariance matrix of $(\theta - \hat{\theta})$ is preferable? Explain your reasoning.
- 9.8** Using the reasons given at the end of Section 9.2.1, explain why SEM is more computationally stable than simply numerically differentiating $\ell(\theta | Y_{(0)})$ twice.

10

Bayes and Multiple Imputation

10.1 Bayesian Iterative Simulation Methods

10.1.1 Data Augmentation

A useful alternative approach to maximum likelihood (ML), particularly when the sample size is small, is to include a reasonable prior distribution for the parameters and compute the posterior distribution of the parameters of interest. We have already been introduced to this approach in Section 6.1.4 with complete data, and with missingness in Sections 7.3 and 7.4.4, in the special case of multivariate normal data with a monotone missingness pattern.

The posterior distribution for a model with ignorable missingness is

$$p(\theta | Y_{(0)}, M) \equiv p(\theta | Y_{(0)}) = \text{const.} \times p(\theta) \times f(Y_{(0)} | \theta), \quad (10.1)$$

where $p(\theta)$ is the prior distribution and $f(Y_{(0)} | \theta)$ is the density of the observed data $Y_{(0)}$. In the examples of Chapter 7, simulation from the posterior distribution could be accomplished without iteration. Specifically, the likelihood was factored into complete data components,

$$L(\phi | Y_{(0)}) = \prod_{q=1}^Q L_q(\phi_q | Y_{(0)}),$$

and, assuming that the parameters (ϕ_1, \dots, ϕ_Q) were also *a priori* independent, the posterior distribution was factored in an analogous way, with ϕ_1, \dots, ϕ_Q a posteriori independent. Consequently, draws $\phi^{(d)} = (\phi_1^{(d)}, \dots, \phi_Q^{(d)})$ of (ϕ_1, \dots, ϕ_Q) could be obtained directly from the factored complete-data posterior distribution. Draws of θ were then obtained as $\theta^{(d)} = \theta(\phi^{(d)})$, where $\theta(\phi)$ is the inverse transformation from ϕ to θ . With more general patterns of missing data or parameters ϕ_j that are not *a priori* independent, this method does not work, and Bayes simulation then requires iteration.

Data augmentation (DA, Tanner and Wong 1987)¹ is an iterative method of simulating the posterior distribution of θ that combines features of the expectation–maximization (EM) algorithm and multiple imputation. It can be thought of as a small-sample refinement of the EM algorithm using simulation, with the imputation (or I) step corresponding to the E step and the posterior (or P) step corresponding to the M step. Start with an initial draw $\theta^{(0)}$ from an approximation to the posterior distribution of θ . Given a value $\theta^{(t)}$ of θ drawn at iteration t :

- (I Step) Draw $Y_{(1)}^{(t+1)}$ with density $p(Y_{(1)} | Y_{(0)}, \theta^{(t)})$.
 (P Step) Draw $\theta^{(t+1)}$ with density $p(\theta | Y_{(0)}, Y_{(1)}^{(t+1)})$.

The procedure is motivated by the fact that the distributions in these two steps are often much easier to draw from than either of the posterior distributions $p(Y_{(1)} | Y_{(0)})$ and $p(\theta | Y_{(0)})$, or the joint posterior distribution $p(\theta, Y_{(1)} | Y_{(0)})$. The iterative procedure can be shown eventually to yield a draw from the joint posterior distribution of $Y_{(1)}, \theta$ given $Y_{(0)}$, in the sense that as t tends to infinity, this sequence converges to a draw from the joint distribution of $(\theta, Y_{(1)})$ given $Y_{(0)}$.

Example 10.1 *Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missingness (Example 8.3 Continued).* Example 8.3 described the EM algorithm for a bivariate normal sample, with one group of units having Y_1 observed but Y_2 missing, a second group of units having both Y_1 and Y_2 observed, and the third group of units having Y_2 observed but Y_1 missing (see Figure 8.1). We now consider DA for this example.

Each iteration t consists of an I step and a P step. The I step of DA is similar to the E step, except that each missing value is replaced by a draw from its conditional distribution given the observed data and the current values of the parameters, rather than by its conditional mean. Because units are conditionally independent given the parameters, each missing y_{i2} is drawn independently as

$$y_{i2}^{(t+1)} \sim_{\text{ind}} N\left(\beta_{20 \cdot 1}^{(t)} + \beta_{21 \cdot 1}^{(t)} y_{i1}, \sigma_{22 \cdot 1}^{(t)}\right),$$

where $\beta_{20 \cdot 1}^{(t)}$, $\beta_{21 \cdot 1}^{(t)}$, and $\sigma_{22 \cdot 1}^{(t)}$ are the t th iterates of the regression parameters of Y_2 on Y_1 . Analogously, each missing y_{i1} is drawn independently as follows:

$$y_{i1}^{(t+1)} \sim_{\text{ind}} N\left(\beta_{10 \cdot 2}^{(t)} + \beta_{12 \cdot 2}^{(t)} y_{i2}, \sigma_{11 \cdot 2}^{(t)}\right),$$

where $\beta_{10 \cdot 2}^{(t)}$, $\beta_{12 \cdot 2}^{(t)}$, and $\sigma_{11 \cdot 2}^{(t)}$ are the t th iterates of the regression parameters of Y_1 on Y_2 .

In the P step of DA, these drawn values of the missing data are treated as if they were the actual observed values of the data, and one draw of the bivariate normal parameters is made from the complete-data posterior distribution, given in Example 6.21. In the limit, the draws are from the joint posterior distribution of the missing data and the parameters. Thus one run of data augmentation generates both a draw from the posterior predictive distribution of $Y_{(1)}$ and a draw from the posterior distribution of θ . Data augmentation can be run independently D times to generate D independent, identically distributed (iid) draws from the joint posterior distribution of θ and $Y_{(1)}$. The values of $Y_{(1)}$ are multiple imputations of the missing values, drawn from their joint posterior predictive distribution.

Note that unlike EM, estimates of the sampling covariance matrix from the filled-in data can be computed without any corrections to the estimated variances. The reason is that draws from the predictive distribution of the missing values are imputed in the I step of DA, rather than the conditional means as in the E step of EM. The loss of efficiency from imputing draws is limited when the posterior mean from DA is computed by averaging over many draws from the posterior distribution, and hence over many imputed data sets.

Example 10.2 Bayesian Computations for a One-Parameter Multinomial Model (Example 9.1 Continued). Example 8.2 applied EM and Example 9.1 applied supplemented expectation–maximization (SEM) to the one-parameter multinomial model of Example 8.2; slightly different asymptotic approximations underlie the calculations in the raw and logit scales. With DA, this distinction is avoided, although different prior distributions yield different posterior distributions. The I step of DA imputes y_3 and $y_4 = 125 - y_3$ assuming the drawn value of θ , $\theta^{(t)}$, is true. Specifically, the I step for iteration $(t + 1)$ of DA draws

$$y_3^{(t+1)} \sim \text{Bin}(125, \theta^{(t)} / (\theta^{(t)} + 2)),$$

which is analogous to the E step of EM given by Eq. (8.10). The complete-data likelihood is proportional to

$$(1/2 - \theta/2)^{y_1} (\theta/4)^{y_2} (\theta/4)^{y_3} (1/2)^{y_4}.$$

Hence with a Beta (Dirichlet) prior distribution proportional to $\theta^{\alpha_1-1} (1-\theta)^{\alpha_2-1}$, the complete-data posterior distribution of θ is a Beta distribution, with density proportional to

$$\theta^{y_2+y_3+\alpha_1-1} (1-\theta)^{y_1+\alpha_2-1}.$$

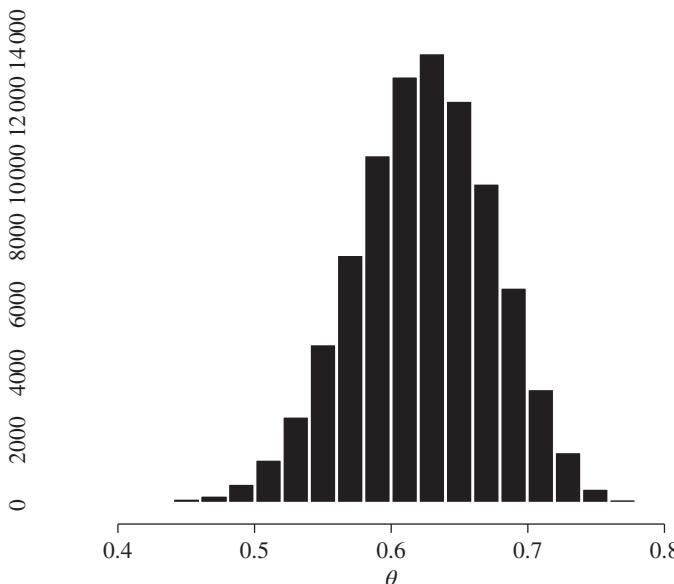


Figure 10.1 Posterior distribution of θ (Jeffreys' prior).

The P step of DA draws from this Beta distribution, with y_1 , y_2 , and y_3 fixed at their values from the previous I step, that is,

$$\theta^{(t+1)} \sim \text{Beta}(y_2 + y_3^{(t+1)} + \alpha_1, y_1 + \alpha_2),$$

using gamma or chi-squared deviates, as described in Example 6.21. This P step is analogous to the M step of EM, Eq. (8.11).

A histogram of 90 000 draws from the posterior distribution of θ , with the Jeffreys' prior distribution with $\alpha_1 = \alpha_2 = 0.5$, is displayed in Figure 10.1. This posterior distribution looks quite close to normal, although the posterior distribution of $\text{logit}(\theta)$ (not shown here) looks more normal. Table 10.1 summarizes the estimated posterior means and standard deviations of θ and $\text{logit}(\theta)$ from this analysis, and the analysis based on the uniform prior distribution for θ , $\alpha_1 = \alpha_2 = 1$. These are close to the ML estimates and asymptotic standard errors from EM/SEM, displayed in the last row of Table 10.1.

10.1.2 The Gibbs' Sampler

The Gibbs' sampler is an iterative simulation method that is designed to yield a draw from the joint distribution in the case of a general pattern of missingness and provides a Bayesian method analogous to the ECM algorithm for ML estimation. In some ways, the Gibbs' sampler is simpler to understand than ECM, because all of its steps involve draws of random variables.

Table 10.1 Example 10.2, estimates from Bayesian and maximum likelihood analyses of multinomial example

Method of analysis	Post. mean/ML estimate of θ	Post. Std. Dev./asymptotic SE of θ	Post. mean/ML estimate of logit θ	Post. Std. Dev./asymptotic SE of logit θ
Bayes, Jeffreys' prior	0.624	0.0515	0.513	0.222
Bayes, uniform prior	0.623	0.0508	0.508	0.219
Maximum likelihood	0.626	0.0515	0.519	0.220

The Gibbs' sampler eventually generates a draw from the distribution $p(x_1, \dots, x_J)$ of a set of J random variables X_1, \dots, X_J in settings where draws from the joint distribution are hard to compute, but draws from conditional distributions $p(x_j | x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_J)$, $j = 1, \dots, J$ are relatively easy to compute. Initial values $x_1^{(0)}, \dots, x_J^{(0)}$ are chosen in some way. Then given values $x_1^{(t)}, \dots, x_J^{(t)}$ at iteration t , new values are found by drawing from the following sequence of J conditional distributions:

$$\begin{aligned} x_1^{(t+1)} &\sim p\left(x_1 | x_2^{(t)}, x_3^{(t)}, \dots, x_J^{(t)}\right), \\ x_2^{(t+1)} &\sim p\left(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_J^{(t)}\right), \\ x_3^{(t+1)} &\sim p\left(x_3 | x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \dots, x_J^{(t)}\right), \\ &\vdots \\ x_J^{(t+1)} &\sim p\left(x_J | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{J-1}^{(t+1)}\right). \end{aligned}$$

It can be shown that, under quite general conditions, the sequence of iterates $x^{(t)} = (x_1^{(t)}, \dots, x_J^{(t)})$ converges to a draw from the joint distribution of X_1, \dots, X_J . In this method, the individual components X_j can be sets of variables, not just scalar variables.

When $J = 2$, the Gibbs' sampler is essentially the same as data augmentation if $X_1 = Y_{(1)}$, $X_2 = \theta$, and distributions condition on $Y_{(0)}$. Then we can, in the limit, obtain a draw from the joint distribution of $(Y_{(1)}, \theta | Y_{(0)})$ by applying the Gibbs' sampler, where at iteration t for the d th imputed data set:

$$Y_{(1)}^{(d,t+1)} \sim p(Y_{(1)} | Y_{(0)}, \theta^{(d,t)}); \quad \theta^{(d,t+1)} \sim p\left(\theta | Y_{(1)}^{(d,t+1)}, Y_{(0)}\right).$$

As with DA, one run of Gibbs' converges to a draw from the posterior predictive distribution of $Y_{(1)}$ and a draw from the posterior distribution of θ . The Gibbs' sampler can be run independently D times to generate D iid draws from the approximate joint posterior distribution of θ and $Y_{(1)}$. The values of $Y_{(1)}$ are multiple imputations of the missing values, drawn from their posterior predictive distribution. The Gibbs' sampler can be used in more complex problems where DA is difficult to compute, but partitioning the missing data or the parameters into more than one piece can help computation. These ideas are illustrated by the following important example.

Example 10.3 *A Multivariate Normal Regression Model with Incomplete Data (Example 8.6 Continued).* Suppose we have n independent observations from the following K -variate normal model:

$$y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma), \quad i = 1, \dots, n, \quad (10.2)$$

where X_i is a known $(K \times p)$ design matrix for the i th observation, β is a $(p \times 1)$ vector of unknown regression coefficients, and Σ is a $(K \times K)$ unknown unstructured variance–covariance matrix. Example 8.6 discussed ML estimation for this problem. We assume the following Jeffreys' prior for the parameters $\theta = (\beta, \Sigma)$:

$$p(\beta, \Sigma) \propto |\Sigma|^{-(K+1)/2}.$$

Draws from the posterior distribution of θ can be obtained from the Gibbs sampler, applied in three steps consisting of an imputation step (I) for $Y_{(1)}$ and two conditional posterior steps (CP1 and CP2) for drawing the values of β and Σ . Let $(Y_{(1)}^{(d,t)}, \beta^{(d,t)}, \Sigma^{(d,t)})$ denote draws of the missing data and parameters after iteration t for creating multiple imputation d . The $(t+1)$ th iteration then consists of the following three steps:

I step: The conditional distribution of $Y_{(1)}$, given $Y_{(0)}$, $\beta^{(d,t)}$, and $\Sigma^{(d,t)}$ is multivariate normal. Let $y_{(0)i}$ and $y_{(1)i}$ denote the sets of observed and missing values in observation i , respectively. Then $y_{(1)i}$ given $Y_{(0)}$, $\beta^{(d,t)}$, and $\Sigma^{(d,t)}$ are independent over i , and multivariate normal with mean and residual covariance matrix based on the linear regression of $y_{(1)i}$ on $y_{(0)i}$ and X_i . Draws $y_{(1)i}^{(d,t+1)}$ from this distribution are readily accomplished using the sweep operator, as discussed in detail in Section 11.2.

CP1 step: The conditional distribution of β , given $Y_{(0)}$, $Y_{(1)}^{(d,t+1)}$, and $\Sigma^{(d,t)}$ is normal with mean

$$\hat{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^n X_i^T (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1} \left\{ \sum_{i=1}^n X_i^T (\Sigma^{(d,t)})^{-1} y_i^{(d,t+1)} \right\}, \quad (10.3)$$

where $y_i^{(d,t+1)} = (y_{(0)i}, y_{(1)i}^{(d,t+1)})$, and covariance matrix

$$\Sigma_{\beta}^{(d,t+1)} = \left\{ \sum_{i=1}^n X_i^T (\Sigma^{(d,t)})^{-1} X_i \right\}^{-1}.$$

Hence $\beta^{(d,t+1)}$ is a random draw from this multivariate normal distribution.

CP2 step: The conditional distribution of Σ given $Y_{(0)}, Y_{(1)}^{(d,t+1)}$ and $\beta^{(d,t+1)}$ is inverse Wishart with scale matrix given by the sum of squares and cross-products matrix of the residuals:

$$\Sigma^{(t+1)} = n^{-1} \sum_{i=1}^n \left(y_i^{(d,t+1)} - X_i \beta^{(d,t+1)} \right) \left(y_i^{(d,t+1)} - X_i \beta^{(d,t+1)} \right)^T \quad (10.4)$$

and degrees of freedom n .

Example 10.4 Univariate t Sample with Known Degrees of Freedom (Example 8.10 Continued). In Example 8.10, we applied the PX-EM algorithm to compute ML estimates for the univariate t model (8.22) with known degrees of freedom v , by imbedding the observed data X in a larger data set (X, W) from the expanded complete-data model:

$$(x_i | \mu_*, \sigma_*, \alpha, w_i) \sim_{\text{ind}} N(\mu_*, \sigma_*^2 / w_i), \quad (w_i | \mu_*, \sigma_*, \alpha) \sim_{\text{ind}} \alpha \chi_v^2 / v, \quad (10.5)$$

with parameters $\phi = (\mu_*, \sigma_*, \alpha)$. This model reduces to the original model (8.23) when $\alpha = 1$. Applying DA to this expanded model yields the Bayesian analog to PX-EM, which is called parameter-expanded data augmentation (PX-DA). The steps of PX-DA in this example are as follows:

The PX-I step is analogous to the PX-E step (8.42) of PX-EM: at iteration $(t+1)$, draw the “missing data” w_i conditionally given x_i and the current draw of parameters $\phi^{(t)}$. From the E step in Example 8.10, this distribution is

$$w_i^{(t)} \sim_{\text{ind}} \chi_{v+1}^2 / (v + d_i^{(t)2}), \quad (10.6)$$

where $d_i^{(t)} = \sqrt{\alpha^{(t)}}(x_i - \mu_*^{(t)})/\sigma_*^{(t)} = (x_i - \mu^{(t)})/\sigma^{(t)}$, as in (8.42).

The PX-M step maximizes the expected complete-data log-likelihood of the expanded model with respect to ϕ . The PX-P step of PX-DA draws ϕ from its complete-data posterior distribution, which is normal-inverse chi-squared as described in Example 6.16.

In Chapter 12, we generalize this PX-DA algorithm to provide a form of robust Bayes inference for multivariate data with missing values.

10.1.3 Assessing Convergence of Iterative Simulations

If the DA or Gibbs' sampler iterations have not proceeded long enough, the simulations may be seriously unrepresentative of the target distribution. Assessing convergence of the sequence of draws to the target distribution is more difficult than assessing convergence of an EM-type algorithm to the ML estimate, because there is no single target quantity to monitor like the maximum value of the likelihood. Methods have been proposed for assessing convergence of a single sequence (see, for example, Geyer 1992, and discussion). However, these methods are only recommended for well-understood models and straightforward datasets. A more reliable approach is to simulate $D > 1$ sequences with starting values dispersed throughout the parameter space. The convergence of all quantities of interest can then be monitored by comparing variation between and within simulated sequences, until the “within” variation roughly equals the “between” variation. Only when the distribution of each simulated sequence is close to the distribution of all the sequences mixed together can they all be approximating the target distribution.

Gelman and Rubin (1992) develop an explicit monitoring statistic based on this idea. For each scalar estimand ψ , label the draws from D parallel sequences as $\psi_{d,t}$ ($d = 1, \dots, D, t = 1, \dots, T$), and compute B and \bar{V} , the between and within sequence variances:

$$B = \frac{T}{D-1} \sum_{d=1}^D (\bar{\psi}_{d\cdot} - \bar{\psi}_{..})^2, \quad \text{where } \bar{\psi}_{d\cdot} = \frac{1}{T} \sum_{t=1}^T \psi_{d,t}, \bar{\psi}_{..} = \frac{1}{D} \sum_{d=1}^D \bar{\psi}_d,$$

$$\bar{V} = \frac{1}{D} \sum_{d=1}^D s_d^2, \quad \text{where } s_d^2 = \frac{1}{T-1} \sum_{t=1}^T (\psi_{d,t} - \bar{\psi}_{d\cdot})^2.$$

We can estimate $\text{Var}(\psi | Y_{\text{obs}})$, the marginal posterior variance of the estimand, by a weighted average of \bar{V} and B , namely

$$\widehat{\text{Var}}^+(\psi | Y_{(0)}) = \frac{T-1}{T} \bar{V} + \frac{1}{T} B,$$

which *overestimates* the marginal posterior variance assuming the starting distribution is appropriately over-dispersed, but is *unbiased* under stationarity (that is, if the starting distribution equals the target distribution). This is analogous to the classical variance estimate for cluster sampling. For any finite T , the “within” variance \bar{V} should be an *underestimate* of $\text{Var}(\psi | Y_{(0)})$ because individual sequences have not had time to range over all the target distribution, and, as a result, should have smaller variance than B ; in the limit as $T \rightarrow \infty$, the expectation of \bar{V} approaches $\text{Var}(\psi | Y_{(0)})$. These facts suggest monitoring convergence of the iterative simulation by estimating the factor by which the

scale of the current distribution for ψ might be reduced if the simulations were continued in the limit as $T \rightarrow \infty$. This potential scale reduction is estimated by

$$\sqrt{\hat{R}} = \sqrt{\widehat{\text{Var}}^+(\psi | Y_{(0)}) / \bar{V}},$$

which declines to 1 as $T \rightarrow \infty$. If the potential scale reduction is high, then there is evidence that proceeding with further simulations should improve our inference about the target distribution. Thus, if $\sqrt{\hat{R}}$ is not near one for all the estimands of interest, the simulation runs should be continued, or perhaps the simulation algorithm itself should be altered to make the simulations more efficient. Once $\sqrt{\hat{R}}$ is near 1 for all scalar estimands of interest, subsequent draws from all the multiple sequences should be collected and treated as draws from the target distribution. The precise implementation of the condition that $\sqrt{\hat{R}}$ is “near” 1 depends on the problem at hand; for most examples, values below 1.2 are acceptable, but for an important analysis or dataset, a higher level of precision may be required.

It is useful to monitor convergence by computing $\sqrt{\hat{R}}$ for the logarithm of the posterior density, as well as for particular quantities of interest. When monitoring scalar quantities of interest, it is best to transform them to be approximately normal (for example, by taking logarithms of all-positive quantities or logits of quantities that lie between 0 and 1). Note that simulation inference from one run with correlated draws is generally less precise than from the same number of independent draws, because of serial correlation within the run. If the simulation efficiency is unacceptably low (in the sense of taking too long to obtain approximate convergence of posterior inference for quantities of interest), seek ways to alter the algorithm to speed convergence (Liu and Rubin 1996, 2002; Gelman et al. 2013).

10.1.4 Some Other Simulation Methods

When draws from the sequence of conditional distributions that form a Gibbs’ algorithm are not easily computed, other simulation approaches are needed. Drawing from complicated multivariate distributions is a very rapidly developing field of statistics (Liu 2001), with many applications outside what might be considered “missing data” problems. However, a variety of the methods have their roots in the missing-data formulation, such as sequential imputation in computational biology (Kong et al. 1994; Liu and Chen 1998). Here we give a brief overview of some of the main ideas, with references.

Suppose that draws of θ are sought from a target distribution $f(\theta)$, but are hard to compute. However, draws are easily obtained from an approximation to the target distribution, say $g(\theta)$, with the same support as $f(\theta)$, and both $f(\theta)$ and $g(\theta)$ can be evaluated up to some proportionality constant. For example, in the

context of Bayesian inference, $f(\theta)$ may be the posterior distribution of logistic regression coefficients, and $g(\theta)$ could be its large-sample normal approximation. A helpful idea involves the use of importance weights to improve the draws from $g(\theta)$, so they can be used as approximate draws from $f(\theta)$. Suppose D^* draws $\theta_1^*, \dots, \theta_{D^*}^*$ are made from $g(\theta)$, where $D^* \gg D$, the desired number of draws from $f(\theta)$, and let $R_d \propto f(\theta_d)/g(\theta_d)$. If D values of θ are drawn from the D^* draws $\theta_1^*, \dots, \theta_{D^*}^*$ with probability proportional to the “importance” ratios or weights, R_d , then in the limit as $D/D^* \rightarrow 0$, the resulting D draws will be from $f(\theta)$.

This simple use of importance weights is known as sampling importance resampling (SIR, see Rubin 1987b; Gelfand and Smith 1990; Smith and Gelfand 1992). More sophisticated uses of these weights involve sequentially accepting or rejecting the draws depending on whether R_d is greater than or less than some constant (rejection sampling, attributed to Von Neumann 1951), or embedding rejection sampling within a Gibbs sampler (the Metropolis–Hastings algorithm, see Metropolis et al. 1953; Hastings 1970). The Gibbs’ sampler and more complex extensions such as the Metropolis–Hastings algorithm are often referred to generically as “Markov Chain Monte Carlo” (MCMC) algorithms, because the sequence of iterates $\theta_{d,1}, \theta_{d,2}, \dots$ forms a Markov chain. Gelman et al. (2013, chapter 11) provide details.

The idea of using the draws from an incorrect distribution to build a “bridge” to the target distribution is the central idea behind bridge sampling, discussed in Meng and Wong (1996). An extension builds a “path” of distributions between the drawing distribution and the target distribution (Gelman and Meng 1998).

Another approach for obtaining approximate draws from a target distribution is to create a set of initial independent parallel draws from a MCMC sequence and analyze them well before they have had a chance to converge to the target distribution. Assuming approximate normality of the target distribution, this estimation is straightforward (Liu and Rubin 1996, 2002) and can be used to create a dramatically improved starting distribution. This “Markov normal” analysis may also reveal subspaces in which the proposed MCMC method is hopelessly slow to converge, and where alternative methods must be used.

10.2 Multiple Imputation

10.2.1 Large-Sample Bayesian Approximations of the Posterior Mean and Variance Based on a Small Number of Draws

The iterative simulation methods we have discussed eventually create draws from the posterior distribution of θ . If inferences for θ are based on the

empirical distribution of the draws (for example a 95% posterior interval of a parameter based on the 2.5 and 97.5 percentiles of the empirical distribution of that parameter), then a large number of independent draws is required, say in the thousands. If, on the other hand, we can assume approximate normality of the observed-data posterior distribution, we need only enough draws to estimate reliably the mean and large sample variance of the posterior distribution, say a few hundred or even fewer. Intermediate numbers of draws might suffice to estimate the posterior distribution by smoothing the empirical distribution, for example, by fitting a parametric model such as the t family, or by semiparametric methods.

In situations with a small fraction of missing information and where inference from the complete-data posterior distribution is based on the multivariate normal or t distribution, posterior moments of θ can be reliably estimated from a surprisingly small number, D , of draws of the missing data $Y_{(1)}$ (e.g., $D = 5–10$). This approach creates D draws of $(\theta, Y_{(1)})$ and applies the combining rules for multiple imputation introduced in Section 5.4.

The idea, first proposed in Rubin (1978b), is to relate the observed-data posterior distribution (10.1) to the “complete-data” posterior distribution that would have been obtained if we had observed the missing data $Y_{(1)}$, namely:

$$p(\theta | Y_{(0)}, Y_{(1)}) \propto p(\theta)L(\theta | Y_{(0)}, Y_{(1)}). \quad (10.7)$$

Equations (10.1) and (10.7) can be related by standard probability theory as follows:

$$p(\theta | Y_{(0)}) = \int p(\theta, Y_{(1)} | Y_{(0)})dY_{(1)} = \int p(\theta | Y_{(1)}, Y_{(0)})p(Y_{(1)} | Y_{(0)})dY_{(1)}. \quad (10.8)$$

Equations (10.8) implies that the posterior distribution of θ , $p(\theta | Y_{(0)})$, can be simulated by first drawing the missing values, $Y_{(1)}^{(d)}$, from their joint posterior distribution, $p(Y_{(1)} | Y_{(0)})$, imputing the drawn values to complete the dataset, and then drawing θ from its “completed-data” posterior distribution, $p(\theta | Y_{(0)}, Y_{(1)}^{(d)})$. When posterior means and variances are adequate summaries of the posterior distribution, (10.8) can be effectively replaced by

$$E(\theta | Y_{(0)}) = E[E(\theta | Y_{(1)}, Y_{(0)}) | Y_{(0)}], \quad (10.9)$$

and

$$\text{Var}(\theta | Y_{(0)}) = E[\text{Var}(\theta | Y_{(1)}, Y_{(0)}) | Y_{(0)}] + \text{Var}[E(\theta | Y_{(1)}, Y_{(0)}) | Y_{(0)}]. \quad (10.10)$$

Multiple imputation effectively approximates the integral (10.8) over the missing values as the average:

$$p(\theta | Y_{(0)}) \approx \frac{1}{D} \sum_{d=1}^D p(\theta | Y_{(1)}^{(d)}, Y_{(0)}), \quad (10.11)$$

where $Y_{(1)}^{(d)} \sim p(Y_{(1)} | Y_{(0)})$ are draws of Y_{mis} from the posterior predictive distribution of the missing values.

Similarly, the mean and variance equations (10.9) and (10.10) can be approximated using the simulated values of $Y_{(1)}$ as follows:

$$E(\theta | Y_{(0)}) \approx \int \theta \frac{1}{D} \sum_{d=1}^D p(\theta | Y_{(1)}^{(d)}, Y_{(0)}) d\theta = \bar{\theta}, \quad (10.12)$$

where $\bar{\theta} = \sum_{d=1}^D \hat{\theta}_d / D$, and $\hat{\theta}_d = E(\theta | Y_{(1)}^{(d)}, Y_{(0)})$ is the estimate of θ from the d th completed data set, and for scalar θ :

$$\text{Var}(\theta | Y_{(0)}) \approx \frac{1}{D} \sum_{d=1}^D V_d + \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta})^2 = \bar{V} + B, \quad (10.13)$$

say, where V_d is the complete-data posterior variance of θ calculated for the d th data set $(Y_{(1)}^{(d)}, Y_{(0)})$, $\bar{V} = \sum_{d=1}^D V_d / D$ is the average of V_d over the D multiply-imputed data sets, and $B = \sum_{d=1}^D (\hat{\theta}_d - \bar{\theta})^2 / (D-1)$ is the between-imputation variance. When D is small, the posterior mean is still approximated by (10.12), but an improved approximation for the posterior variance (10.13) is obtained by multiplying the between-imputation component by $(1+D^{-1})$, that is

$$\text{Var}(\theta | Y_{(0)}) \approx \bar{V} + (1+D^{-1})B. \quad (10.14)$$

The ratio of estimated between-imputation to total variance, $\hat{\gamma}_D = (1+D^{-1})B / (\bar{V} + (1+D^{-1})B)$, estimates the fraction of missing information. For vector θ , the variance V_d is replaced by a covariance matrix, and $(\hat{\theta}_d - \bar{\theta})^2$ is replaced by $(\hat{\theta}_d - \bar{\theta})(\hat{\theta}_d - \bar{\theta})^T$.

A further refinement for small D is to replace the normal reference distribution by a t distribution with degrees of freedom given by

$$\nu = (D-1) \left(1 + \frac{D}{D+1} \frac{\bar{V}}{B} \right)^2. \quad (10.15)$$

When the completed data sets are based on limited degrees of freedom, say v_{com} , an additional refinement replaces v with

$$v^* = (v^{-1} + \hat{v}_{\text{obs}}^{-1})^{-1}, \quad \text{where } \hat{v}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{v_{\text{com}} + 1}{v_{\text{com}} + 3} \right) v_{\text{com}}. \quad (10.16)$$

The theoretical bases for (10.15) and (10.16) are given in Rubin and Schenker (1986), Rubin (1987a), and Barnard and Rubin (1999).

Example 10.5 Bivariate Normal Data with Ignorable Nonresponse and a General Pattern of Missingness (Example 10.1 Continued). Suppose that the algorithm of Example 10.1 is run independently five times to create five joint draws of θ and Y_{mis} . Five draws are far too few to generate a reliable empirical distribution for estimating the actual posterior distribution of θ . However, the five draws of $Y_{(1)}$ can be adequate for generating MI inferences based on the methods of this section, provided the fraction of missing information is modest, as when the fractions of units with Y_1 or Y_2 missing are limited. In that case, the draws of $Y_{(1)}$ yield five “completed” datasets, the d th with sample means, variances, and covariance that we denote $\{(\bar{y}_1^{(d)}, \bar{y}_2^{(d)}, s_{11}^{(d)}, s_{22}^{(d)}, s_{12}^{(d)}), d = 1, \dots, 5\}$. The resulting estimate of μ_1 from Eq. (10.12) is

$$\tilde{\mu}_1 = \sum_{d=1}^5 \bar{y}_1^{(d)} / 5,$$

with associated standard error from Eq. (10.14),

$$\text{Var}(\mu_1) = (1/5) \sum_{d=1}^5 \left(s_{11}^{(d)} / n \right) + (6/5)(1/4) \sum_{d=1}^5 \left(\bar{y}_1^{(d)} - \tilde{\mu}_1 \right)^2.$$

If the original sample size n is large, a 95% interval estimate of μ_1 is given by

$$\tilde{\mu}_1 \pm t_{v, 0.975} \sqrt{\text{Var}(\mu_1)},$$

where v is given by Eq. (10.15) with $D = 5$. For small n , the more refined approximation (10.16) should be used.

10.2.2 Approximations Using Test Statistics or p -Values

In addition to interval estimation, it is often of interest to summarize the posterior distribution for a multicomponent estimand by calculating a test statistic with an associated p -value. Some multivariate analogs of the expressions given for scalar quantities are listed in Rubin (2004, section 3.4). Meng and Rubin (1992) developed methods for likelihood-ratio testing when the available information consists of point estimates and the function for the evaluation of

the complete-data loglikelihood ratio statistic as a function of these estimates and the completed data. With large data sets and large models, such as in the common situation of a multiway contingency table, the complete-data analysis may produce only a test statistic or p -value, and no parameter estimates. With such limited information, Rubin (1987a, section 3.5) provided initial methods and Li et al. (1991a) developed improved methods that require only the D completed-data chi-squared statistics (or equivalently, the D completed-data p -values) that result from testing a null hypothesis using each of the D completed data sets. These methods, however, are less accurate than methods that use the completed-data statistics $\hat{\theta}_d, V_d$. More recent literature on improved methods (e.g., Harel 2009; Chaurasia and Harel 2015) has loose theoretical motivation. Hence, we start with a summary of the more accurate methods.

For θ with $k > 1$ components, significance levels for null values of θ can be obtained from D completed-data estimates, $\hat{\theta}_d, d = 1, \dots, D$, and their associated large-sample variance–covariance matrices, $V_d, d = 1, \dots, D$, using multivariate analogs of the previous expressions. First, let θ_0 be the null value of θ , and let

$$W(\theta_0, \bar{\theta}) = (\theta_0 - \bar{\theta})^T \bar{V}^{-1} (\theta_0 - \bar{\theta}) / ((1 + r)k), \quad (10.17)$$

where $r = (1 + D^{-1}) \text{ trace}(\bar{V}^{-1})/k$, where $\text{trace}(\bar{V}^{-1})/k$ is the average diagonal element of \bar{V}^{-1} . Equation (10.17) is an estimated Wald statistic, as defined in Section 6.1.3. The p -value is then

$$\Pr(F_{k,\ell} > W(\theta_0, \bar{\theta})), \quad (10.18)$$

where $F_{k,\ell}$ is an F random variable with k and ℓ degrees of freedom with

$$\ell = 4 + (k(D - 1) - 4)(1 + a/r)^2, \quad a = (1 - 2/(k(D - 1))); \quad (10.19)$$

if $k(D - 1) \leq 4$, let $\ell = (k + 1)v/2$. Rubin (2004) and Li et al. (1991b) provide motivation for this test statistic and its reference distribution.

With large models, each complete-data analysis may not produce the complete-data variance–covariance matrix V_d , but a p -value for $\theta = \theta_0$ may still be desired. Two general methods are available, one asymptotically as precise as $W(\theta_0, \bar{\theta})$, and one less precise, but simpler to use. We describe the more accurate method first.

Typically, in multiparameter problems, in addition to the parameter of interest θ , there will be nuisance parameters ϕ , which are estimated by different values when $\theta = \theta_0$ and when $\theta \neq \theta_0$. Let $\hat{\phi}$ be the complete-data estimate of ϕ when $\theta = \hat{\theta}$ and $\hat{\phi}_0$ be the complete-data estimate of ϕ when $\theta = \theta_0$. Assume

the complete-data analysis produces the estimates $(\hat{\theta}, \hat{\phi})$, the null estimates $(\theta_0, \hat{\phi}_0)$ and the p -value for $\theta = \theta_0$ based on the likelihood-ratio χ^2 statistic,

$$p\text{-value} = \Pr(\chi_k^2 > \text{LR}), \quad (10.20)$$

where $\text{LR} = \text{LR}((\hat{\theta}, \hat{\phi}), (\theta_0, \hat{\phi}_0))$, using the notation of Section 6.1.3, and χ_k^2 is a χ^2 random variable on k degrees of freedom. Let the average values of $\hat{\theta}$, $\hat{\phi}$, $\hat{\phi}_0$ and LR across the D sets of multiple imputations be denoted $\bar{\theta}$, $\bar{\phi}$, $\bar{\phi}_0$ and $\bar{\text{LR}}$. Assume that the function LR can be evaluated for each of the D completed data sets at $\bar{\theta}$, $\bar{\phi}$, θ_0 , $\bar{\phi}_0$ to obtain D values of $\text{LR}((\bar{\theta}, \bar{\phi}), (\theta_0, \bar{\phi}_0))$ whose average across the D imputations is $\bar{\text{LR}}_0$. Then

$$\bar{\text{LR}}_0/(k + (D + 1)(\bar{\text{LR}} - \bar{\text{LR}}_0)/(D - 1)) \quad (10.21)$$

is identical in large samples to $W(\theta_0, \bar{\theta})$ and can be used exactly as if it were $W(\theta_0, \bar{\theta})$ (Meng and Rubin 1992).

In some situations, the complete-data method of analysis may not produce estimates of the general function $\text{LR}(\cdot, \cdot)$, but only the value of the likelihood ratio statistic so that the D multiple imputations result in D values $\text{LR}_1, \dots, \text{LR}_D$. If so, the following procedure due to Li et al. (1991a) can be used. Let the repeated-imputation p -value be $\Pr(F_{k,b} > \widetilde{\text{LR}})$, where

$$\widetilde{\text{LR}} = \frac{(\bar{\text{LR}}/k) - (1 - D^{-1})v}{1 + (1 + D^{-1})v}, \quad (10.22)$$

and v is the sample variance of $(\text{LR}_1^{1/2}, \dots, \text{LR}_D^{1/2})$, and

$$b = k^{-3/D}(D - 1)(1 + ((1 + D^{-1})v)^{-1})^2. \quad (10.23)$$

The general method defined by (10.22) and (10.23) can be quite inaccurate in general. With one-sided tests for an associated scalar estimand, the following procedure² can be very accurate.

First consider a one-sided test of a null hypothesis H_0 with complete data. The asymptotic sampling distribution of the one-sided p -value is uniformly distributed on $(0, 1)$. Therefore, the inverse normal cumulative distribution transformation of the p -value p , say $z = \Phi^{-1}(p)$, has an asymptotic standard normal distribution under H_0 , with negative values of z corresponding to values of the scalar test statistic smaller than expected under H_0 , and positive values of z corresponding to values of the scalar test statistic larger than expected under H_0 .

Now suppose that instead of complete data, the data set is incomplete, and the missing values have been multiply imputed, creating D completed data sets. Applying the procedure in the previous paragraph to each completed data set

yields D p -values $\{p_d, d = 1, \dots, D\}$ and corresponding standard normal deviates $\{z_d, d = 1, \dots, D\}$. If the imputed data were actually the observed data, then under H_0 they would be standard normal, and hence have sampling variance 1.0. Thus, the set $\{z_d\}$ can be combined using the MI combining rules, yielding a combined test statistic equal to

$$z_{\text{MI}} = \frac{\bar{z}}{\sqrt{1 + (1 + 1/D)s_z^2}}, \text{ where } \bar{z} = \sum_{d=1}^D z_d / D \text{ and } s_z^2 = \sum_{d=1}^D (z_d - \bar{z})^2 / (D - 1). \quad (10.24)$$

The combined one-sided test of H_0 then compares z_{MI} with a standard normal distribution.

10.2.3 Other Methods for Creating Multiple Imputations

We now return to the problem of creating multiple imputations. The theory of the previous section suggests that we draw

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} \mid Y_{(0)}) \quad (10.25)$$

from the posterior predictive distribution of the missing values. Unfortunately, it is often difficult to draw from this predictive distribution in complicated problems, because of the implicit requirement in (10.25) to integrate over the parameters θ . Data augmentation accomplishes this by iteratively drawing a sequence of values of the parameters and missing data until convergence. Although this approach is theoretically preferable if the underlying model is well justified, in situations with multivariate data involving nonlinear relationships, building one coherent model for the joint distribution of the variables, programming the draws, and assessing convergence may be difficult and time-consuming. Simpler methods that approximate draws from Eq. (10.25), although less formally rigorous, may be easier to implement and yield approximately valid inferences when used in conjunction with the combining rules in Sections 10.2.1 and 10.2.2. Such methods may even be more effective than rigorous MI inference under a full model, if the full model is not a good reflection of the process that generated the data.

A trivial example of an approximate method is to run the simulation for a fixed number of iterations or fixed time, without formally assessing convergence. We now describe some alternatives:

1. *Improper MI*: An approximate method is to draw:

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} \mid Y_{(0)}, \tilde{\theta}), \quad (10.26)$$

where $\tilde{\theta}$ is an estimate of θ , for example the ML estimate, or an easy-to-compute estimate such as that from the complete units. This is a reasonable approximation with small fractions of missing information, but Rubin (1987a, Chapter 4) shows that it does not provide valid frequentist inferences in general, because uncertainty in estimating θ is not propagated. Rubin (1987a) calls methods that do not propagate this uncertainty *improper*.

2. *Use the posterior distribution from a subset of the data:* Often, it is relatively simple to draw θ from its posterior distribution based on a subset of the data close to the full data. The method propagates uncertainty about θ , but does not use all the available information to draw θ . For example, we have seen in Chapter 7 that the posterior distribution of θ may have a simple form for a monotone missing data pattern. This insight suggests discarding values to create a data set $Y_{(0)\text{mp}}$ with a monotone pattern, and then drawing θ from its posterior distribution given $Y_{(0)\text{mp}}$. That is, draw $Y_{(1)}^{(d)}$ as follows:

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} \mid Y_{(0)}, \tilde{\theta}^{(d)}), \quad \text{where } \tilde{\theta}^{(d)} \sim p(\theta \mid Y_{(0)\text{mp}}). \quad (10.27)$$

An even simpler but less accurate example of this approach is to draw θ from its posterior distribution given the complete units, that is,

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} \mid Y_{(0)}, \tilde{\theta}^{(d)}), \quad \text{where } \tilde{\theta}^{(d)} \sim p(\theta \mid Y_{(0)\text{cc}}), \quad (10.28)$$

where $Y_{(0)\text{cc}}$ represents data from the complete units. For the multivariate normal problem with missing values (10.28) can be viewed as a stochastic version of Buck's method (see Example 4.3), and is related to a class of pattern-mixture models involving complete-unit missing value restrictions, as discussed in Little (1993c).

3. *Filling in data to create a monotone pattern:* In some situations, where a monotone missing-data pattern is destroyed by a small number of missing values, an attractive option is to impute these “nonmonotone” missing values using one of the single imputation methods of Chapter 4, preferably as draws from an approximation to their posterior predictive distribution:

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} \mid Y_{(0)}, \tilde{\theta}^{(d)}), \quad \text{where } \tilde{\theta}^{(d)} \sim p(\theta \mid Y_{\text{aug-mp}}),$$

where $Y_{\text{aug-mp}}$ is the observed data augmented to create a monotone pattern. This method could be combined with method 2, using subsets of the data, in various ways.

4. *Use the asymptotic distribution of the ML estimate:* Suppose the ML estimate $\hat{\theta}$ of θ is available, together with an estimate of its large-sample covariance

matrix $C(\hat{\theta})$, as discussed in Section 6.1.2. Then $\theta^{(d)}$ can be drawn from its asymptotic normal posterior distribution:

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} | Y_{(0)}, \tilde{\theta}^{(d)}), \quad \text{where } \tilde{\theta}^{(d)} \sim N[\hat{\theta}, C(\hat{\theta})].$$

The draw $\theta^{(d)}$ has the form $\theta^{(d)} = \hat{\theta} + z^{(d)}$, where $z^{(d)}$ is multivariate normal with mean 0 and covariance matrix $C(\hat{\theta})$. In large samples, this method is clearly preferable to method 1 and often preferable to method 2, because it correctly propagates asymptotic uncertainty in the ML estimate of θ .

5. *Refining approximate draws using importance sampling:* Methods 2–4 draw pairs $(Y_{(1)}^{(d)}, \tilde{\theta}^{(d)})$ from a joint distribution where the draw of $Y_{(1)}^{(d)}$ given $\tilde{\theta}^{(d)}$ is correct but the draw of $\tilde{\theta}^{(d)}$ is from an approximating density, say $g(\theta)$. A refinement is obtained by drawing a substantial set (for example, 100–1000) of draws $Y_{(1)}^{(d)}$, and then subsampling a smaller number (for example, 2–10) from this set, with probability of the selection of draw d proportional to $w_d \propto p(\tilde{\theta}^{(d)})L(\tilde{\theta}^{(d)} | Y_{(0)})/g(\tilde{\theta}^{(d)})$. This is a version of SIR (see Section 10.1.4) and was used in the large-sale application described in Rubin (1983b). As the ratio of the initial set to the final number of draws gets large, the final draws are correct under mild support conditions.
6. *Substituting ML estimates from bootstrapped samples:* If EM is used to estimate θ and the large-sample covariance matrix is not readily available, then an approximate draw from the posterior distribution can be obtained as the estimate from applying EM to a bootstrapped sample $Y_{(0)}^{(\text{boot},d)}$ of the observed data, that is, a random sample with replacement from $Y_{(0)}$ of the same size as $Y_{(0)}$. That is,

$$Y_{(1)}^{(d)} \sim p(Y_{(1)} | Y_{(0)}, \tilde{\theta}^{(d)}), \quad \text{where } \tilde{\theta}^{(d)} = \hat{\theta}(Y_{(0)}^{(\text{boot},d)}).$$

This procedure is asymptotically proper, in the sense that the ML estimates from the bootstrap samples are asymptotically equivalent to a sample from the posterior distribution of θ . This method may provide some robustness to model misspecification because the bootstrap provides estimates of uncertainty asymptotically equivalent to the sandwich estimator (6.17). However, if a substantial fraction of the bootstrap samples do not yield unique ML estimates and are discarded, the standard errors based on the remaining samples can be severely underestimated.

7. *Predictive mean matching multiple imputation:* In Chapter 4, we described the hot deck, a single imputation method that matches each unit with missing values (called the recipient unit) to a unit with complete data (called the donor unit) based on closeness with respect to a metric. The missing values for the recipient are then imputed using the corresponding values from the

donor. These methods can be extended to MI, by creating a set of donors for each incomplete unit, and then imputing missing values for each completed data set from a randomly selected unit from the donor set (Little 1988c). The method can reflect uncertainty in the process for creating the donor sets by creating a bootstrap sample of the donor set prior to imputation of each set of missing values.

The predictive-mean matching metric Eq. (4.10) in Example 4.11 measures closeness in terms of the predictive distribution of the missing values. An issue here is how many units to include in the donor set – a small number increases the closeness of the matches, but a large number provides a larger choice of donor values for imputation. Even small sets of donors provide for some propagation of imputation uncertainty, a key objective of the MI method.

This method has been implemented as an option in a number of software packages, including SAS PROC MI. A potential advantage of the method over parametric model-based MI is that the imputations are less vulnerable to model misspecification, because the model is only used to create a metric, rather than being used directly to generate the predictive distribution of the imputations. A disadvantage is that the method relies on the quality of the matching of donors to recipients, and some donors may be relatively far from recipients with respect to the predictive mean metric, leading to potential bias. These properties suggest that predictive mean matching MI is more suited to large data sets where donors are relatively plentiful, and bias from misspecification of the model underlying the matching metric is more important than precision, rather than small data sets where donors are sparse and precision is the predominant concern. The results of the simulation study in Schenker and Taylor (1996) support the propriety of statement.

10.2.4 Chained-Equation Multiple Imputation

MI is described in detail in Chapters 11–14 for a variety of models for the joint distribution of the variables, including the multivariate normal for continuous variables, loglinear models for categorical variables, and the general location model for mixtures of continuous and categorical variables. A disadvantage of these methods is that these standard joint distributions do not always provide a good fit to real multivariate data. On the other hand, it is often possible to formulate a set of conditional distributions relating each variable to the collection of other variables, which appear reasonable when taken one at a time, but are incoherent in the sense that they cannot be derived from a single joint distribution. Such models, even when incoherent, may be useful for creating multiple imputations Chained-Equation Multiple Imputation. Chained-Equation Multiple Imputation is based on this idea, and creates imputations based on a set

of conditional distributions, using a method analogous to the Gibbs' sampler described in Section 10.1.2.

Specifically, let X_1, \dots, X_K be a set of variables with missing values, Z a vector of variables that are fully observed, $x_{(0)}$ be the set of observed data on X_1, \dots, X_K , and $x_{j(1)}$ the set of missing values for $X_j, j = 1, \dots, K$. For $j = 1, \dots, K$, we specify a suitable model for the conditional distribution of X_j given $(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_K, Z)$ with density $p_j(x_j | x_{(0)}, \dots, x_{j-1}, x_{j+1}, \dots, x_K, z, \theta_j)$, parameters θ_j , and prior distribution $\pi_j(\theta_j)$. The models take into account the nature of outcome variable – for example, logistic or probit for a binary variable – and can include nonlinear terms and interactions of the predictors, if appropriate. They do not, however, necessarily correspond to a coherent joint distribution for X_1, \dots, X_K . Missing values are imputed via the following steps:

- (a) Create initial imputations $x_{1(1)}^{(0)}, \dots, x_{K(1)}^{(0)}$ of the missing values by some approximate procedure.
- (b) Given current imputed values $x_{1(1)}^{(t)}, \dots, x_{K(1)}^{(t)}$ at iteration t , generate updated imputed values for each variable as draws from the following sequence of p predictive distributions:

$$\begin{aligned} x_{1(1)}^{(t+1)} &\sim p\left(x_{1(1)} | x_{(0)}, z, x_{2(1)}^{(t)}, x_{3(1)}^{(t)}, \dots, x_{K-1(1)}^{(t)}, x_{K(1)}^{(t)}\right), \\ &\vdots \\ x_{j(1)}^{(t+1)} &\sim p\left(x_{j(1)} | x_{(0)}, z, x_{1(1)}^{(t+1)}, x_{j-1(1)}^{(t+1)}, x_{j(1)}^{(t)}, \dots, x_{K(1)}^{(t)}\right), \\ &\vdots \\ x_{K(1)}^{(t+1)} &\sim p\left(x_{K(1)} | x_{(0)}, z, x_{1(1)}^{(t+1)}, x_{2(1)}^{(t+1)}, x_{3(1)}^{(t+1)}, \dots, x_{K-1(1)}^{(t)}\right). \end{aligned}$$

In particular, the set of draws $\{x_{j(1)}^{(t+1)}\}$ in this sequence are created by first drawing $\theta_j^{(t+1)}$ from the posterior distribution of θ_j , given $(x_{(0)}, z, x_{1(1)}^{(t+1)}, \dots, x_{j-1(1)}^{(t+1)}, x_{j(1)}^{(t)}, \dots, x_{K(1)}^{(t)})$, and then drawing $\{x_{j(1)}^{(t+1)}\}$ from their posterior predictive distribution, given $(x_{(0)}, z, x_{1(1)}^{(t+1)}, \dots, x_{j-1(1)}^{(t+1)}, x_{j(1)}^{(t)}, \dots, x_{K(1)}^{(t)})$, and $\theta_j^{(t+1)}$; for iid models, the latter draws are independent over the units, and are typically computationally straightforward.

When the set of conditional distributions correspond to a coherent joint distribution, this algorithm is a Gibbs' sampler, and the sequence converges to draws from the correct posterior distribution. In particular, Bayes and MI for the bivariate normal model in Example 10.1 can be implemented by the chained-equations approach, with each conditional distribution being normal with a linear additive regression on the conditioned variables, with a constant residual variance. In other settings, convergence to a stationary distribution is

not guaranteed; however, the procedure appears to produce useful imputations, provided the conditional distributions yield good fits to the observed data. The increased flexibility in modeling these conditional distributions may outweigh the lack of clear theoretical justification of the method.

Kennickell (1991) is an early application of the chained equations method to an important survey; see also the discussion in Rubin (2017). Software implementing the method includes MICE (Van Buuren and Oudshoorn 1999); and IVEWARE (Raghunathan et al. 2001). Rubin (2002) proposes limiting the possibly incoherent draws to missing values that need to be filled in to create a monotone pattern, and then imputing the remaining missing values coherently via a sequence of conditional distributions appropriate for the artificially created monotone missingness pattern. Again, the method lacks theoretical support, but it seems to work well in at least some applications (Li et al. 2014).

Because chained-equations multiple imputations are based on a sequence of regression models, it is important to conduct diagnostic checks to ensure that the imputations are reasonable. One possibly useful approach is to estimate the missingness propensity for each variable with missing values, conditional on the observed or imputed values of other variables, and then compare the distribution of the observed values with the distribution of one set of the imputed values, within categories of the estimated missingness propensity. If the imputation model is generating reasonable imputations, these empirical distributions should look similar. For more discussion of graphical and diagnostic checks (see Aboyomi et al. 2008; Bondarenko and Raghunathan 2016).

10.2.5 Using Different Models for Imputation and Analysis

If the entire rationale for doing multiple imputation were for the computation of Bayesian posterior distributions in large samples, it would be an important but relatively limited tool. As the examples in Section 10.2.3 suggest, however, often a method can be chosen for multiple imputation without consideration of the precise method to be used for the analysis of the multiply-imputed data. If the model underlying the method chosen to impute the data and the model chosen for analysis are identical, the theory is as described in Section 10.2.1. A theoretically interesting and practically important setting occurs when the imputation method does *not* perfectly align with the complete-data analysis conducted by the ultimate user. That is, the ultimate user of the multiply-imputed data could apply a variety of simple or potentially complicated complete data analyses to the multiply-imputed data, and then use the combining rules and combined results without reference to how the imputations were created.

Somewhat surprisingly, this approach can be successful, especially with relatively limited fractions of missing information, as suggested by theoretical

results and as documented by empirical examples. A simple example illustrates this phenomenon.

Example 10.6 *Inference Under the Approximate Bayesian Bootstrap (Example 5.8 Continued).* Suppose that the approximate Bayesian bootstrap (ABB) method of Example 5.8 is used to create multiple imputations within adjustment cells, but that the complete-data analysis will be based on the large-sample normality of the sample mean. Assuming MAAR, it is simple to show that the combining rules give valid frequentist inferences. In fact, this result holds for a variety of other multiple-imputation methods: fully normal, the Bayesian bootstrap, a mean and variance-adjusted hot-deck, etc. (see examples 4.1–4.4 in Rubin 2004).

When the imputation method uses more information than the complete data analysis, and this information is correct, the complete-data analyses of the multiply-imputed data will tend to be more efficient than anticipated: for instance confidence intervals will have greater than the nominal coverage. This phenomenon was noted in Rubin and Schenker (1987) and Fay (1992, 1996), and termed “super-efficiency” in Rubin (1996).

The general situation is called “uncongeniality” of the imputer’s and ultimate user’s models by Meng (1995). Usually, uncongeniality leads to conservative inferences, although in special circumstances it can lead to invalid (i.e., anti-conservative) inferences. The following example conveys some intuition; other examples are discussed by Meng (2002), Robins and Wang (2000), Xie and Meng (2017, with discussion).

Example 10.7 *Effects of a Misspecified Imputation Model.* Suppose we have a sample of units (X, Y) , where X is fully observed, but half the values of Y are missing due to an MAR process. In truth, Y is a monotone but a nonlinear function of X , $Y = \exp(X)$. Suppose multiple imputations of missing Y values are created using a linear model relating Y to X . Clearly, the residual variability of Y on X will be overestimated due to lack of fit; the true residual variability is zero, and if an exponential model were fit, this would be found. The extra residual variability in the linear model has two consequences for multiple imputation of the missing Y s. First, the between-imputation variability (e.g., in the slope of the linear model for Y on X) will be greater than if the true model were fit, and second, for each set of imputations, the individual imputations (on and off the regression line) will be more variable than if the correct model were used. Thus, both between and within variability are overestimated relative to their values if the correct model were applied. Because the linear fit often gives a decent approximation to the truth for global estimands, such as the grand mean or median, using an incorrect model for multiple imputation typically leads to overestimated variability, for such global estimands and thus, overcoverage of their interval estimates. This result is seen in simulations with real data (e.g.,

Raghunathan and Rubin 1998). With estimands in the tails of the distribution such as quartiles, this approximate validity does not generally hold.

In our limited experience with real and artificial data sets (e.g., Ezzati-Rice et al. 1995), the practical conclusion appears to be that multiple imputation, when carefully done, can be safely used with real problems even when the ultimate user may be applying models or analyses not contemplated by the imputer.

Problems

- 10.1** Reproduce the posterior distribution in Figure 10.1, and compare the posterior mean and standard deviation with that given in Table 10.1. Recalculate the posterior distribution of θ using the improper prior distribution with $\alpha_1 = \alpha_2 = 0$. Is the resulting posterior distribution proper?
- 10.2** Consider a simple random sample of size n sampled from a finite population of size N , with r respondents and $m = n - r$ nonrespondents, and let \bar{y}_R and s_R^2 be the sample mean and variance of the respondents' data, and \bar{y}_{NR} and s_{NR}^2 the sample mean and variance of the imputed data. Show that the mean \bar{y}_* and variance s_*^2 of all the data can be written as
- $$\begin{aligned}\bar{y}_* &= (r\bar{y}_R + m\bar{y}_{NR})/n \text{ and } s_*^2 = ((r-1)s_R^2 + (m-1)s_{NR}^2 \\ &\quad + rm(\bar{y}_R - \bar{y}_{NR})^2/n)/(n-1).\end{aligned}$$
- 10.3** Suppose in Problem 10.2 that imputations are randomly drawn with replacement from the r respondents' values. Assume the missing data are MACAR.
- (a) Show that \bar{y}_* is unbiased for the population mean \bar{Y} .
 - (b) Show that conditional on the observed data, the sampling variance of \bar{y}_* is $ms_R^2(1-r^{-1})/n^2$, and that the expectation of s_*^2 is $s_R^2(1-r^{-1})(1+rn^{-1}(n-1)^{-1})$.
 - (c) Show that conditional on the sample sizes n and r (and the population Y -values), the sampling variance of \bar{y}_* is the variance of \bar{y}_R times $(1+(r-1)n^{-1}(1-r/n)(1-r/N)^{-1})$, and show that this is greater than the expectation of $U_* = s_*^2(n^{-1} - N^{-1})$.
 - (d) Assume r and N/r are large, and show that interval estimates of \bar{Y} based on using U_* as the estimated sampling variance of \bar{y}_* are too short by a factor $(1+nr^{-1}-rn^{-1})^{1/2}$. Note that there are two reasons: $n > r$, and \bar{y}_* is not as efficient as \bar{y}_R . Tabulate true coverages and true significance levels as functions of r/n and nominal level.

- 10.4** Suppose D multiple imputations are created using the method of Problem 10.3, and let $\bar{y}_*^{(d)}$ and $U_*^{(d)}$ be the values of \bar{y}_* and U_* for the d th completed dataset. Let $\bar{\bar{y}}_* = \sum_{d=1}^D \bar{y}_*^{(d)}/D$, and T_* be the multiple-imputation estimate of the sampling variance of the $\bar{\bar{y}}_*$. That is

$$T_* = \bar{U}_* + (1 + D^{-1})B_*,$$

$$\text{where } \bar{U}_* = \sum_{d=1}^D U_*^{(d)}/D, B_* = \sum_{d=1}^D (\bar{y}_*^{(d)} - \bar{\bar{y}}_*)^2/(B - 1).$$

- (a) Show that, conditional on the data, the expected value of B_* equals the variance of the $\bar{y}_*^{(d)}$.
 - (b) Show that the variance of $\bar{\bar{y}}_*$ (conditional on n, r , and the population Y -values) is $D^{-1} \text{Var}(\bar{y}_*) + (1 - D^{-1}) \text{Var}(\bar{y}_R)$, and conclude that $\bar{\bar{y}}_*$ is more efficient than the single-imputation estimate \bar{y}_* .
 - (c) Tabulate values of the relative efficiency of $\bar{\bar{y}}_*$ to \bar{y}_R for different values of D , assuming large r and large N/r .
 - (d) Show that the sampling variance of $\bar{\bar{y}}_*$ (conditional on n, r , and the population Y -values) is greater than the expectation of T_* by approximately $s_R^2(1 - r/n)^2/r$.
 - (e) Assume r and N/r are large, and tabulate true coverages and significance levels of the multiple-imputation inference. Compare with the results in Problem 10.3, part (d).
- 10.5** Modify the multiple-imputation approach of Problem 10.4 to give the correct answer for large r and N/r . (*Hint:* For example, add a random residual $s_R r^{-1/2} z_d$ to the imputed value for unit i .)

Notes

- 1 The definition of DA used here differs slightly from the original version, which involves a multiple imputation step at each iteration, followed by multiple draws of the parameters from the current estimate of the posterior distribution.
- 2 This procedure is not published in a journal, but was proposed and used by Rubin in a 2008 US Food and Drug Administration submission for Emphasys Medical's Zephyr, an endobronchial valve device.

Part III

Likelihood-Based Approaches to the Analysis of Incomplete Data: Some Examples

11

Multivariate Normal Examples, Ignoring the Missingness Mechanism

11.1 Introduction

In this chapter, we apply the tools of Part II to a variety of common problems involving incomplete data on multivariate normally distributed variables: estimation of the mean vector and covariance matrix; estimation of these quantities when there are restrictions on the mean and covariance matrix; multiple linear regression, including analysis of variance (ANOVA), and multivariate regression; repeated measures models, including random coefficient regression models where the coefficients themselves are regarded for maximum likelihood (ML) computations as missing data; and selected time series models. Robust estimation with missing data is discussed in Chapter 12, the analysis of partially-observed categorical data is considered in Chapter 13, and the analysis of mixed continuous and categorical data is considered in Chapter 14. Chapter 15 concerns models with data missing not at random.

11.2 Inference for a Mean Vector and Covariance Matrix with Missing Data Under Normality

Many multivariate statistical analyses, including multiple linear regression, principal components analysis, discriminant analysis, and canonical correlation analysis, are based on the initial summary of the data matrix into the sample mean and covariance matrix of the variables. Thus inference for the population mean and covariance matrix for an arbitrary pattern of missing values is a particularly important problem. In Sections 11.2.1 and 11.2.2 we discuss ML for the mean and covariance matrix from an incomplete multivariate normal sample, assuming the missingness mechanism is ignorable. Section 11.2.3 describes Bayesian inference and multiple imputation (MI) for this problem. Although

the assumption of multivariate normality may appear restrictive, the methods discussed here can provide consistent estimates under weaker assumptions about the underlying distribution. The multivariate normality assumption will be relaxed somewhat when we consider linear regression in Section 11.4 and robust estimation in Chapter 12.

11.2.1 The EM Algorithm for Incomplete Multivariate Normal Samples

Suppose that (Y_1, Y_2, \dots, Y_K) have a K -variate normal distribution with mean $\mu = (\mu_1, \dots, \mu_K)$ and covariance matrix $\Sigma = (\sigma_{jk})$. We write $Y = (Y_{(0)}, Y_{(1)})$, where Y represents a random sample of size n on (Y_1, \dots, Y_K) , $Y_{(0)}$ the set of observed values, and $Y_{(1)}$ the missing data. Also, let $y_{(0),i}$ represent the set of variables with values observed for unit i , $i = 1, \dots, n$. The loglikelihood based on the observed data is then

$$\ell(\mu, \Sigma | Y_{(0)}) = \text{const} - \frac{1}{2} \sum_{i=1}^n \ln |\Sigma_{(0),i}| - \frac{1}{2} \sum_{i=1}^n (y_{(0),i} - \mu_{(0),i})^T \Sigma_{(0),i}^{-1} (y_{(0),i} - \mu_{(0),i}), \quad (11.1)$$

where $\mu_{(0),i}$ and $\Sigma_{(0),i}$ are the mean and covariance matrix of the observed components of Y for unit i .

To derive the expectation–maximization (EM) algorithm for maximizing (11.1), we note that the hypothetical complete data Y belong to the regular exponential family (8.19) with sufficient statistics

$$S = \left(\sum_{i=1}^n y_{ij}, j = 1, \dots, K; \quad \sum_{i=1}^n y_{ij} y_{ik}, j, k = 1, \dots, K \right).$$

At the t th iteration of EM, let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ denote current estimates of the parameters. The E step of the algorithm for iteration $t + 1$ calculates

$$E \left(\sum_{i=1}^n y_{ij} \mid Y_{(0)}, \theta^{(t)} \right) = \sum_{i=1}^n y_{ij}^{(t+1)}, \quad j = 1, \dots, K \quad (11.2)$$

and

$$E \left(\sum_{i=1}^n y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)} \right) = \sum_{i=1}^n \left(y_{ij}^{(t+1)} y_{ik}^{(t+1)} + c_{jki}^{(t+1)} \right), \quad j, k = 1, \dots, K, \quad (11.3)$$

where

$$y_{ij}^{(t+1)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed,} \\ E(y_{ij} \mid y_{(0),i}, \theta^{(t)}), & \text{if } y_{ij} \text{ is missing,} \end{cases} \quad (11.4)$$

and

$$c_{jki}^{(t+1)} = \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed,} \\ \text{Cov}(y_{ij}, y_{ik} | y_{(0),i}, \theta^{(t)}), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing.} \end{cases} \quad (11.5)$$

Missing values y_{ij} are thus replaced by the conditional mean of y_{ij} given the set of values, $y_{(0),i}$ observed for that unit and the current estimates of the parameters, $\theta^{(t)}$. These conditional means and the nonzero conditional covariances are easily found from the current parameter estimates by sweeping the augmented covariance matrix so that the variables $y_{(0),i}$ are predictors in the regression equation and the remaining variables $y_{(1),i}$ are outcome variables. The sweep operator is described in Section 7.4.3. Note that Eqs. (11.2) and (11.4) impute the best linear predictors of the missing values given current estimates of the parameters, thus showing the link between ML and efficient imputation of the missing values. Equation (11.3) includes adjustments c_{jki} needed to correct for biases in the resulting estimated covariance matrix from imputing conditional means for the missing values.

The M step of the EM algorithm is straightforward. The new estimates $\theta^{(t+1)}$ of the parameters are computed from the estimated complete-data sufficient statistics. That is,

$$\begin{aligned} \mu_j^{(t+1)} &= n^{-1} \sum_{i=1}^n y_{ij}^{(t+1)}, \quad j = 1, \dots, K; \\ \sigma_{jk}^{(t+1)} &= n^{-1} E \left(\sum_{i=1}^n y_{ij} y_{ik} | Y_{(0)}, \theta^{(t)} \right) - \mu_j^{(t+1)} \mu_k^{(t+1)} \\ &= n^{-1} \sum_{i=1}^n \left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)} \right) \left(y_{ik}^{(t+1)} - \mu_k^{(t+1)} \right) + c_{jki}^{(t+1)} \right], \quad j, k = 1, \dots, K. \end{aligned} \quad (11.6)$$

Beale and Little (1975) suggest replacing the factor n^{-1} in the estimate of σ_{jk} by $(n-1)^{-1}$, which parallels the correction for degrees of freedom in the complete-data case.

It remains to suggest initial values of the parameters. Four straightforward possibilities are (i) to use the complete-case solution of Section 3.2; (ii) to use one of the available-case (AC) solutions of Section 3.4; (iii) to form the sample mean and covariance matrix of the data filled in by one of the single-imputation methods of Chapter 4; or (iv) to form means and variances from observed values of each variable and set all starting correlations equal to zero. Option (i) provides consistent estimates of the parameters if the data are missing completely at random (MCAR) and there are at least $K+1$ complete observations. Option (ii) makes use of all the available data but can yield an estimated covariance matrix that is not positive definite, leading to possible problems in the first iteration. Options (iii) and (iv) generally yield inconsistent estimates

of the covariance matrix, but estimates that are either positive semidefinite (Option iii) or positive definite (Option iv), and hence are usually workable as starting values. A computer program for general use should have several alternative initializations of the parameters available so that a suitable choice can be made. Another reason for having a variety of starting values available is to examine the likelihood for multiple maxima.

Orchard and Woodbury (1972) first described this EM algorithm. Earlier, the scoring algorithm for this problem had been described by Trawinski and Bargmann (1964) and Hartley and Hocking (1971). An important difference between scoring and EM is that the former algorithm requires inversion of the information matrix of μ and Σ at each iteration. After convergence, this matrix provides an estimate of the asymptotic covariance matrix of the ML estimates, which is not needed by, nor obtained by, the EM computations. The inversion of the information matrix of θ at each iteration, however, can be expensive because this is a large matrix if the number of variables is large. For the K -variable case, the information matrix of θ has $K + K(K+1)/2$ rows and columns, and when $K = 30$ it has over 100 000 elements. With EM, an asymptotic covariance matrix of θ can be obtained by supplemented expectation–maximization (SEM), bootstrapping, or by just one inversion of the information matrix evaluated at the final ML estimate of θ , as described in Chapter 9.

Three versions of EM can be defined. The first stores the raw data (Beale and Little 1975). The second stores the sums, sums of squares, and sums of cross products for each pattern of missing data (Dempster et al. 1977). Because the version that takes less storage and computation is to be preferred, a preferable option is a third, which mixes the two previous versions, storing raw data for those patterns with fewer than $(K+1)/2$ units and storing sufficient statistics for the other more frequent patterns.

11.2.2 Estimated Asymptotic Covariance Matrix of $(\theta - \hat{\theta})$

Let $\theta = (\mu, \Sigma)$, where Σ is represented as a row vector $(\sigma_{11}, \sigma_{12}, \sigma_{22}, \dots, \sigma_{KK})$. If the data are MCAR, the expected information matrix of θ has the form

$$J(\theta) = \begin{bmatrix} J(\mu) & 0 \\ 0 & J(\Sigma) \end{bmatrix}.$$

Here, the (j, k) th element of $J(\mu)$, corresponding to row μ_j , column μ_k , is

$$\sum_{i=1}^n \psi_{jki},$$

where

$$\psi_{jki} = \begin{cases} (j, k)\text{th element of } \Sigma_{(0),i}^{-1}, & \text{if both } x_{ij} \text{ and } x_{ik} \text{ are observed,} \\ 0, & \text{otherwise,} \end{cases}$$

and $\Sigma_{(0),i}$ is the covariance matrix of the variables observed for unit i . The $(\ell m, rs)$ th element of $J(\Sigma)$, corresponding to row $\sigma_{\ell m}$, column σ_{rs} , is

$$\frac{1}{4}(2 - \delta_{\ell m})(2 - \delta_{rs}) \sum_{i=1}^n (\psi_{\ell ri}\psi_{msi} + \psi_{\ell si}\psi_{mri}),$$

where $\delta_{\ell m} = 1$ if $\ell = m$, 0 if $\ell \neq m$. As noted earlier, the inverse of $J(\hat{\theta})$ supplies an estimated covariance matrix for the ML estimate $\hat{\theta}$. The matrix $J(\theta)$ is estimated and inverted at each step of the scoring algorithm. Note that the expected information matrix is block diagonal with respect to the means and the covariances. Hence, if these asymptotic variances are only required for ML estimates of means or linear combinations of means, then it is only necessary to calculate and invert the information matrix $J(\mu)$ corresponding to the means, which has relatively small dimension.

The observed information matrix, which is calculated and inverted at each iteration of the Newton–Raphson algorithm, is not even block diagonal with respect to μ and Σ , so this complete-data simplification does not occur. On the other hand, the standard errors based on the observed information matrix can be viewed as valid when the data are missing at random (MAR) but not MCAR, and hence, should be preferable to those based on $J(\theta)$ in applications. For more discussion, see Kenward and Molenberghs (1998). As noted above, EM does not yield an information matrix, so if any such matrix is used as a basis for standard errors, it must be calculated and inverted after the ML estimates are obtained, as with SEM described in Section 9.2.1. A simple alternative with sufficient data is to compute the ML estimates on bootstrap samples, and apply the methods of Section 9.2.2.

11.2.3 Bayes Inference and Multiple Imputation for the Normal Model

We now describe a Bayesian analysis of the multivariate normal model in Section 11.2.1. To simplify the description, we assume the conventional Jeffreys' prior distribution for the mean and covariance matrix:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(K+1)/2},$$

and present an iterative data augmentation (DA) algorithm for generating draws from the posterior distribution of $\theta = (\mu, \Sigma)$:

$$p(\mu, \Sigma | Y_{(0)}) \propto |\Sigma|^{-(K+1)/2} \exp(\ell(\mu, \Sigma | Y_{(0)})),$$

where $\ell(\mu, \Sigma | Y_{(0)})$ is the loglikelihood in Eq. (11.1). Let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ and $Y^{(t)} = (Y_{(0)}, Y_{(1)}^{(t)})$ denote current draws of the parameters and filled-in data matrix at iteration t . The I step of the DA algorithm simulates

$$Y_{(1)}^{(t+1)} \sim p(Y_{(1)} | Y_{(0)}, \theta^{(t)}).$$

Because the rows of the data matrix Y are conditionally independent given θ , this is equivalent to drawing

$$y_{(1),i}^{(t+1)} \sim p(y_{(1),i} | y_{(0),i}, \theta^{(t)}) \quad (11.7)$$

independently for $i = 1, \dots, n$. As noted in the discussion of EM, this distribution is multivariate normal with mean given by the linear regression of $y_{(1),i}$ on $y_{(0),i}$, evaluated at current draws $\theta^{(t)}$ of the parameters. The regression parameters and residual covariance matrix of this normal distribution are obtained computationally by sweeping on the augmented covariance matrix

$$\Sigma^{*(t)} = \begin{pmatrix} -1 & \mu^{(t)\top} \\ \mu^{(t)} & \Sigma^{(t)} \end{pmatrix},$$

so that the observed variables are swept in (conditioned on) and the missing variables are swept out (being predicted). The draw $y_{(1),i}^{(t+1)}$ is simply obtained by adding to the conditional mean in the E step of EM, Eqs. (11.2) and (11.4), a normal draw with mean 0 and a function of the current draw of the covariance matrix of the missing variables given the observed variables in unit i , say $\Sigma_{(1)\cdot(0),i}^{(t)}$.

The P step of DA draws

$$\theta^{(t+1)} \sim p(\theta | Y^{(t+1)}),$$

where $Y^{(t+1)} = (Y_{(0)}, Y_{(1)}^{(t+1)})$ is the imputed data from $(t+1)$ st the I step (11.7). The draw of $\theta^{(t+1)}$ can be accomplished in two steps:

$$\begin{aligned} (\Sigma^{(t+1)} / (n - 1) | Y^{(t+1)}) &\sim \text{Inv-Wishart}(S^{(t+1)}, n - 1), \\ (\mu^{(t+1)} | \Sigma^{(t+1)}, Y^{(t+1)}) &\sim N_K(\bar{y}^{(t+1)}, \Sigma^{(t+1)} / n), \end{aligned} \quad (11.8)$$

where $(\bar{y}^{(t+1)}, S^{(t+1)})$ is the sample mean and covariance matrix of Y from the imputed data $Y^{(t+1)}$. The posterior distribution of θ can be simulated directly using Eqs. (11.7) and (11.8), after a suitable burn-in period to achieve stationary draws. For more computational details on the P step, see Example 6.19.

An alternative analysis is MI, which creates sets of draws of the missing data based on Eq. (11.7), and then derives inferences using the MI combining rules given in Section 10.2. The Chained Equation algorithm discussed in Section 10.2.4, with normal linear additive regressions for the conditional distributions of each variable given the others, provides an alternative to the DA algorithm. It also yields draws from the predictive distribution of the missing values that can be used to create MI data sets, although (as usually implemented) the

predictive distributions are slightly different because of different choices of prior distributions for the parameters of the conditional distributions.

Example 11.1 *St. Louis Risk Research Data.* We illustrate these methods using data in Table 11.1 from the St. Louis Risk Research Project. One objective of the project was to evaluate the effects of parental psychological disorders on various aspects of the development of their children. Data on $n = 69$ families with two children were collected. Families were classified according to risk group of the parent (G), a trichotomy defined as follows:

1. ($G = 1$), a normal group of control families from the local community.
2. ($G = 2$), a moderate-risk group where one parent was diagnosed as having secondary schizo-affective or other psychiatric illness or where one parent had a chronic physical illness.
3. ($G = 3$), a high-risk group where one parent had been diagnosed as having schizophrenia or an affective mental disorder.

In this example, we compare data on $K = 4$ continuous variables R_1 , V_1 , R_2 and V_2 by risk group G , where R_c and V_c are standardized reading and verbal comprehension scores for the c th child in a family, $c = 1, 2$. The variable G is always observed, but the outcome variables are missing in a variety of different combinations, as seen in Table 11.1. Analysis of two categorical outcome variables D_1 = number of symptoms for first child (1, low; 2, high) and D_2 = number of symptoms for second child (1, low; 2, high) in Table 11.1 is deferred until Chapter 13.

Table 11.2 displays estimates for the four continuous outcomes in the low-risk group and the combined moderate and high-risk groups. The columns show estimates of the mean, standard error of the mean (sem), and the standard deviation from four methods: AC analysis, ML with sem computed using the bootstrap, DA with estimates and standard errors based on 1000 draws of the posterior distribution, and MI based on 10 MI's and the formulae in Section 10.2. Estimates from DA and MI yield very similar results, as expected, and ML is generally similar. The results from AC analysis are broadly similar, but the estimated means deviate noticeably in some cases, namely V_1 and R_2 for the low risk group, and V_1 and V_2 for the moderate/high risk groups. General conclusions of superiority cannot be inferred without knowing the true estimand values, but the ML, DA, and MI estimates appear to make better use of the observed data.

The Bayesian analysis readily provides inferences for other parameters. For example, substantive interest concerns the comparison of means between risk groups. Figure 11.1 shows plots of the posterior distributions of the differences in means for each of the four outcomes, based on 9000 draws. The posterior distributions appear to be fairly normal. The 95% posterior probability intervals

Table 11.1 Example 11.1, St. Louis risk research data

Low risk ($G = 1$)						Moderate risk ($G = 2$)						High risk ($G = 3$)					
First child			Second child			First child			Second child			First child			Second child		
R_1	V_1	D_1	R_2	V_2	D_2	R_1	V_1	D_1	R_2	V_2	D_2	R_1	V_1	D_1	R_2	V_2	D_2
110	?	?	?	150	1	88	85	2	76	78	?	98	110	?	112	103	2
118	165	1	?	130	2	?	98	?	114	133	?	127	138	1	92	118	1
116	145	2	114	125	?	108	103	2	90	100	2	113	?	?	?	?	?
?	?	?	126	?	?	113	?	2	95	115	2	107	93	?	92	75	?
118	140	1	118	123	?	?	65	?	97	68	2	?	?	1	101	?	2
?	120	?	105	128	?	118	?	2	?	?	2	?	?	?	87	98	2
?	?	?	96	113	?	92	?	2	?	?	?	114	?	2	?	?	2
138	163	1	130	140	?	90	?	1	110	?	2	56	58	2	88	105	1
115	153	1	?	?	?	98	123	?	96	88	?	96	95	1	87	100	2
?	145	2	139	185	2	113	110	?	112	115	?	126	135	2	118	133	?
126	138	1	105	133	1	102	130	?	114	120	?	?	?	?	130	195	?
120	160	?	109	150	?	89	113	2	130	135	?	?	?	?	116	?	2
?	133	?	98	108	?	90	80	2	91	75	2	64	45	2	82	53	2
?	?	?	115	140	2	?	?	?	109	88	2	128	?	2	121	?	2
115	158	2	?	135	1	75	63	1	88	13	1	?	120	1	108	118	?
112	115	2	93	140	?	93	?	I	?	?	?	?	?	?	100	140	2
133	168	1	126	158	2	?	?	?	115	?	2	105	138	1	74	75	1
118	180	1	116	148	?	123	170	1	115	138	2	88	118	?	84	103	?
123	?	1	110	155	1	114	130	2	104	123	2						
100	?	1	101	120	1	?	?	2	113	123	2						
118	138	1	?	110	1	113	?	2	?	?	2						
103	108	?	?	?	?	117	?	?	82	103	2						
121	155	1	?	100	2	122	?	1	114	?	2						
?	?	?	?	?	2	105	?	2	?	?	1						
?	?	?	104	118	1												
?	?	?	87	85	1												
?	?	?	?	63	?												

? denotes missing.

Table 11.2 Example 11.1, means and SD's of continuous outcomes in low-, medium-, and high-risk groups, St. Louis risk research data

Variable	Low risk ($G = 1$)			Moderate- and high-risk ($G = 2, 3$)		
	Mean	sem	SD	Mean	sem	SD
V_1						
AC	146.1	4.8	19.7	105.5	6.6	30.8
ML	143.4	5.5	19.5	115.7	5.9	31.8
DA	143.7	5.4	22.7	115.6	6.3	34.3
MI	143.8	5.4	22.5	115.6	6.3	34.4
V_2						
AC	128.6	5.4	25.9	106.6	5.4	28.9
ML	128.6	5.1	25.7	110.8	5.1	27.8
DA	128.5	6.0	28.6	110.6	5.2	30.0
MI	128.5	6.0	28.6	110.8	5.2	30.0
R_1						
AC	117.9	2.3	9.4	102.7	3.2	17.7
ML	116.8	2.9	10.0	103.4	3.3	18.1
DA	116.8	2.8	12.2	103.4	3.4	19.5
MI	116.8	2.8	12.2	103.3	3.3	19.5
R_2						
AC	110.7	3.2	13.7	101.6	2.5	15.0
ML	108.1	3.0	13.8	101.9	2.5	14.6
DA	108.5	3.4	15.4	101.8	2.7	15.7
MI	108.4	3.4	15.4	101.8	2.7	15.7

Estimates from available-case analysis (AC), maximum likelihood (ML), data augmentation (DA), and multiple imputation (MI), under normal model.

based on the 2.5th–97.5th percentiles are shown below the plots. The fact that three of these four intervals are entirely positive is evidence that reading and verbal means are higher in the low-risk group than in the moderate- and high-risk group.

11.3 The Normal Model with a Restricted Covariance Matrix

In Section 11.2, there were no restrictions on the parameters of the multivariate normal, θ being free to vary anywhere in its natural parameter space. Many

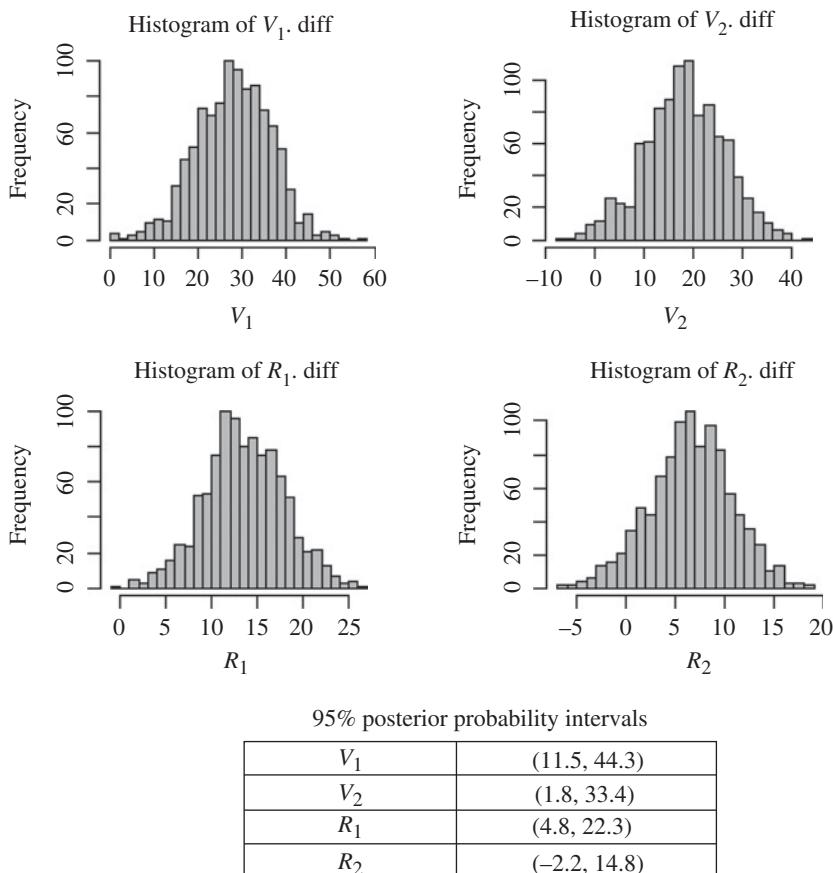


Figure 11.1 Example 11.1, posterior distributions of mean differences $\mu_{\text{low}} - \mu_{\text{med/high}}$, St. Louis risk research data, based on 9000 draws.

important statistical models, however, place restrictions on θ . ML and Bayes for such restricted models with incomplete data can be readily handled, whenever the complete-data analyses subject to the restrictions are simple. The reason is that the E step of EM, or the I step of DA, take the same form whether θ is restricted or not; the only changes are to modify the M step of EM or the P step of DA to be appropriate for the restricted model.

For some kinds of restrictions on θ , noniterative ML or Bayes estimates do not exist even with complete data. In some of these cases, EM or DA can be used to compute ML or Bayes estimates by *creating* fully missing variables in such a way that the M or P step is noniterative. We present EM algorithms for two examples to illustrate this idea. Both examples can be easily modified to handle missing data among the variables with some observed values.

Example 11.2 Patterned Covariance Matrices. Some patterned covariance matrices that do not have explicit ML estimates can be viewed as submatrices of larger patterned covariance matrices that do have explicit ML estimates. In such a case, the smaller covariance matrix, say Σ_{11} , can be viewed as the covariance matrix for observed variables and the larger covariance matrix, say Σ , can be viewed as the covariance matrix for both observed and fully missing variables. In such a case, the EM algorithm can be used to calculate the desired ML estimates for the original problem, as described by Rubin and Szatrowski (1982).

As an illustration, suppose that we have a random sample y_1, \dots, y_n from a multivariate normal distribution, $N_3(0, \Sigma_{11})$, with the 3×3 stationary covariance pattern

$$\Sigma_{11} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_2 & \theta_1 & \theta_2 \\ \theta_3 & \theta_2 & \theta_1 \end{bmatrix}.$$

The ML estimate of Σ_{11} does not have an explicit form. However, these observations can be viewed as the first three of four components from a random sample $(y_1, z_1), \dots, (y_n, z_n)$ from a multivariate normal distribution $N_4(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_2 \\ \theta_2 & \theta_1 & \theta_2 & \theta_3 \\ \theta_3 & \theta_2 & \theta_1 & \theta_2 \\ \theta_2 & \theta_3 & \theta_2 & \theta_1 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

If (y_i, z_i) are all observed, ML estimates of Σ can be computed by simple averaging (Szatrowski 1978). Thus, we apply the EM algorithm, assuming the first three components of each (y_i, z_i) are observed, and the last component, z_i , is missing. The $\{y_i\}$ are the observed data, and the (y_i, z_i) are the complete data, both observed and missing. Let $C = \sum (y_i, z_i)^T (y_i, z_i)/n$ and $C_{11} = \sum (y_i^T y_i)/n$. The matrix C is the complete-data sufficient statistic and C_{11} is the observed sufficient statistic.

There is only one pattern of incomplete data (y_i observed and z_i missing), so the E step of the EM algorithm involves calculating the expected value of C given the observed sufficient statistic C_{11} and the current estimate $\Sigma^{(t)}$ of Σ , namely, $C^{(t)} = E(C | C_{11}, \Sigma^{(t)})$. First, find the regression parameters of the conditional distribution of z_i given y_i by sweeping Y from the current estimate of Σ , $\Sigma^{(t)}$, to obtain

$$\begin{bmatrix} \Sigma_{11}^{(t)-1} & \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} & \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)} \end{bmatrix} = \text{SWP}[1, 2, 3] \begin{bmatrix} \Sigma_{11}^{(t)} & \Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)} & \Sigma_{22}^{(t)} \end{bmatrix}.$$

The expected value of z_i given the observed data and $\Sigma = \Sigma^{(t)}$ is $y_i \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)}$, so that the expected value of $C_{12} = \sum_i y_i^T z_i / n$ given C_{11} and $\Sigma^{(t)}$ is $C_{11} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)}$. The expected value of $z_i^T z_i$ given the observed data and $\Sigma = \Sigma^{(t)}$ is

$$\left(\Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} y_i^T \right) \left(y_i \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)} \right) + \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)},$$

so that the expected value of $C_{22} = \sum_i^n z_i^T z_i / n$ is

$$\Sigma_{22}^{(t)} - \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)} + \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} C_{11} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)}.$$

These calculations are summarized as follows:

$$C^{(t+1)} = E(C | C_{11}, \Sigma^{(t)}) = \begin{bmatrix} C_{11} & C_{11} \Sigma_{11}^{(t)-1} \Sigma_{12}^{(t)} \\ \Sigma_{21}^{(t)} \Sigma_{11}^{(t)-1} C_{11} & \left\{ \Sigma_{22}^{(t)} - \Sigma_{21}^{(t)} \left(\Sigma_{11}^{(t)-1} - \Sigma_{11}^{(t)-1} C_{11} \Sigma_{11}^{(t)-1} \right) \Sigma_{12}^{(t)} \right\} \end{bmatrix}. \quad (11.9)$$

The ML estimate of Σ given the complete data, C , is explicit and as noted above is obtained by simple averaging. Thus, the M step of EM at iteration $t + 1$ is given by

$$\begin{aligned} \theta_1^{(t+1)} &= \frac{1}{4} \left(\sum_{k=1}^4 c_{kk}^{(t+1)} \right), & \theta_2^{(t+1)} &= \frac{1}{4} \left(c_{12}^{(t+1)} + c_{23}^{(t+1)} + c_{34}^{(t+1)} + c_{14}^{(t+1)} \right), \\ \theta_3^{(t+1)} &= \frac{1}{2} \left(c_{13}^{(t+1)} + c_{24}^{(t+1)} \right), \end{aligned} \quad (11.10)$$

where $c_{kj}^{(t+1)}$ is the (k,j) th element of $C^{(t+1)}$, the expected value of C from the E step (11.9) at iteration $t + 1$. These estimates of θ_1 , θ_2 , and θ_3 yield a new estimate of Σ for iteration $t + 1$. This new value of C is used in (11.10) to calculate new estimates of θ_1 , θ_2 , and θ_3 and thus $\Sigma^{(t+1)}$.

An advantage of EM is its ability to handle simultaneously both missing values in the data matrix and patterned covariance matrices, both of which occur frequently in a variety of applications, such as educational testing examples. In some of these examples, unrestricted covariance matrices do not have unique ML estimates because of the missing data, and the patterned structure is easily justified from theoretical considerations and from empirical evidence on related data (Holland and Wightman 1982; Rubin and Szatrowski 1982). When there is more than one pattern of incomplete data, the E step computes expected sufficient statistics for each of the patterns rather than just one pattern as in (11.9).

Example 11.3 Exploratory Factor Analysis. Let Y be an $n \times K$ observed data matrix and Z be an $n \times q$ unobserved “factor-score matrix,” $q < K$, and let (y_i, z_i) denote the i th row of (Y, Z) . Assume

$$\begin{aligned} (y_i | z_i, \theta) &\sim_{\text{ind}} N_K(\mu + \beta z_i, \Sigma), \\ (z_i | \theta) &\sim_{\text{ind}} N_q(0, I_q), \end{aligned} \tag{11.11}$$

where $\beta (K \times q)$ is commonly called the factor-loading matrix, I_q is the $(q \times q)$ identity matrix, $\Sigma = \text{Diag}(\sigma_1^2, \dots, \sigma_K^2)$ is called the uniquenesses matrix, and $\theta = (\mu, \beta, \Sigma)$. Integrating out the unobserved factors z_i yields the exploratory factor-analysis model:

$$(y_i | \theta) \sim_{\text{ind}} N_K(\mu, \beta\beta^T + \Sigma).$$

In factor analysis, it is often assumed that $\mu = 0$; the slightly more general model (11.11) leads to centering the variables Y by subtracting the sample mean for each variable. Little and Rubin (1987) present an EM algorithm for ML estimation of θ . Here we present the faster ML algorithm of Rubin and Thayer (1982), which Liu et al. (1998) show to be an example of a Parameter-expanded expectation–maximization (PX-EM) algorithm.

As discussed in Section 8.5.3, PX-EM creates a model in a larger parameter space where the fraction of missing information is reduced. The expanded model is

$$\begin{aligned} (y_i | z_i, \phi) &\sim_{\text{ind}} N_K(\mu^* + \beta^* z_i, \Sigma^*), \\ (z_i | \phi) &\sim_{\text{ind}} N_q(0, \Gamma), \end{aligned} \tag{11.12}$$

where $\phi = (\mu^*, \beta^*, \Sigma^*, \Gamma)$, and the unrestricted covariance matrix Γ replaces the identity matrix I_q in (11.11). Under model (11.12), $(y_i | \phi) \sim_{\text{ind}} N_K(\mu^*, \beta^*\Gamma\beta^{*\top} + \Sigma^*)$, so

$$\theta = (\mu, \beta, \Sigma) = (\mu^*, \beta^*\text{Chol}(\Gamma), \Sigma^*),$$

where $\text{Chol}(\Gamma)$ is the Cholesky factor of Γ (see Example 6.19). The complete data sufficient statistics for (11.12) (that is, if $\{(y_i, z_i), i = 1, \dots, n\}$ were fully observed), are

$$\begin{aligned} \bar{y} &= \sum_{i=1}^n y_i/n, & C_{yy} &= \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T/n, \\ C_{yz} &= \sum_{i=1}^n (y_i - \bar{y})z_i^T/n, & C_{zz} &= \sum_{i=1}^n z_i z_i^T/n. \end{aligned}$$

Given current parameter estimates $\phi^{(t)}$, the E step of PX-EM consists of computing the expected complete-data sufficient statistics:

$$\begin{aligned} C_{yz}^{(t+1)} &= E(C_{yz} | Y, \phi^{(t)}) = C_{yy}\gamma^{(t)}, \\ C_{zz}^{(t+1)} &= E(C_{zz} | Y, \phi^{(t)}) = \gamma^{(t)\top} C_{yy}\gamma^{(t)} + C_{zz\cdot y}^{(t)}, \end{aligned}$$

where $\gamma^{(t)}$ and $C_{zz\cdot y}^{(t)}$ are the regression coefficients and residual covariance matrix of Z on Y given $\phi^{(t)}$. Specifically, let

$$B^{(t)} = \begin{pmatrix} \beta^{*(t)\top} \beta^{*(t)} + \Sigma^{*(t)} & \beta^{*(t)\top} \\ \beta^{*(t)} & I_q \end{pmatrix}$$

be the current variance–covariance matrix of (Y, Z) ; then $\gamma^{(t)}$ and $C_{zz\cdot y}^{(t)}$ are obtained from the last q columns of $\text{SWP}[1, \dots, K] B^{(t)}$.

The M step of PX-EM calculates the cross-products matrix

$$C^{(t+1)} = \begin{pmatrix} C_{yy} & C_{yz}^{(t+1)} \\ C_{yz}^{(t+1)\top} & C_{zz}^{(t+1)} \end{pmatrix}.$$

It then sets $\mu^{*(t+1)} = \bar{y}$, $\Gamma^{(t+1)} = C_{zz}^{(t+1)}$, and finds $\beta^{*(t+1)}$ and $\Sigma^{*(t+1)}$ from the last q columns of $\text{SWP}[1, \dots, K] C^{(t+1)}$. Reduction to the original parameters θ gives $\mu^{(t+1)} = \mu^{*(t+1)}$, $\Sigma^{(t+1)} = \Sigma^{*(t+1)}$, and $\beta^{(t+1)} = \beta^{*(t+1)} \text{Chol}(\Gamma^{(t+1)})$.

This EM algorithm for factor analysis can be extended to handle missing data $Y_{(1)}$ in the Y variables, by treating both $Y_{(1)}$ and Z as missing data. The E step then calculates the contribution to the expected sufficient statistics from each pattern of incomplete data, rather than just the single pattern with y_i completely observed.

Example 11.4 Variance Component Models. A large collection of patterned covariance matrices arises from variance components models, also called random effects or mixed effects ANOVA models. The EM algorithm can be used to obtain ML estimates of variance components and more generally covariance components (Dempster et al. 1977; Dempster et al. 1981). The following example is taken from Snedecor and Cochran (1967, p. 290).

In a study of artificial insemination of cows, semen samples from $K = 6$ bulls were tested for their ability to conceive, where the number, n_i , of semen samples tested from bulls varied from bull to bull; the data are given in Table 11.3. Interest focuses on the variability of the bull effects; that is, if an infinite number of samples had been taken from each bull, the variance of the six resulting means would be calculated and used to estimate the variance of the bull effects in the

Table 11.3 Data for Example 11.4

Bull (i)	Percentages of conception to services for successive samples	n_i	X_i
1	46, 31, 37, 62, 30	5	206
2	70, 59	2	129
3	52, 44, 57, 40, 67, 64, 70	7	394
4	47, 21, 70, 46, 14	5	198
5	42, 64, 50, 69, 77, 81, 87	7	470
6	35, 68, 59, 38, 57, 76, 57, 29, 60	9	479
Total		35	1876

population. Thus, with the actual data, there is one component of variability due to sampling bulls from a population of bulls, which is of primary interest, and another due to samples from each bull.

A common normal model for such data is

$$y_{ij} = \alpha_i + e_{ij}, \quad (11.13)$$

where $(\alpha_i \mid \theta) \sim_{\text{ind}} N(\mu, \sigma_\alpha^2)$ are the between-bull effects, $(e_{ij} \mid \theta) \sim_{\text{ind}} N(0, \sigma_e^2)$ are the within-bull effects, and $\theta = (\mu, \sigma_\alpha^2, \sigma_e^2)$ are fixed parameters. Integrating over the α_i , the y_{ij} are jointly normal with common mean μ , common variance $\sigma_e^2 + \sigma_\alpha^2$, and covariance σ_α^2 within the same bull and 0 between bulls. That is

$$\text{Corr}(y_{ij}, y_{i'j'}) = \begin{cases} \rho = [1 + \sigma_e^2/\sigma_\alpha^2]^{-1}, & \text{if } i = i', j \neq j', \\ 0, & \text{if } i \neq i', \end{cases}$$

where ρ is commonly called the intraclass correlation.

Treating the unobserved random variables $\alpha_1, \dots, \alpha_6$ as missing data (with all y_{ij} observed) leads to an EM algorithm for obtaining ML estimates of θ . Specifically, the complete-data likelihood has two factors, the first corresponding to the distribution of y_{ij} given α_i and θ , and the second to the distribution of α_i given θ :

$$\prod_{i,j} (2\pi\sigma_e^2)^{-1/2} \exp[-(y_{ij} - \alpha_i)^2 / (2\sigma_e^2)] \prod_i (2\pi\sigma_\alpha^2)^{-1/2} \exp[-(\alpha_i - \mu)^2 / (2\sigma_\alpha^2)].$$

The resulting loglikelihood is linear in the following complete-data sufficient statistics:

$$T_1 = \sum \alpha_i, \quad T_2 = \sum \alpha_i^2, \quad T_3 = \sum_{i,j} (y_{ij} - \bar{\alpha}_i)^2 = \sum_{i,j} (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \alpha_i)^2.$$

The E step of EM takes the expectations of T_1 , T_2 , T_3 given current estimates of θ and the observed data y_{ij} , $i = 1, \dots, K$, $j = 1, \dots, n_i$. These follow by applying Bayes's theorem to the joint distribution of the α_i and the y_{ij} to obtain the conditional distribution of the α_i given the y_{ij} :

$$(\alpha_i | \{y_{ij}\}, \theta) \sim_{\text{ind}} N(w_i \mu + (1 - w_i)\bar{y}_i, v_i),$$

where $w_i = \sigma_\alpha^{-2} v_i$ and $v_i = (\sigma_\alpha^{-2} + n_i \sigma_e^{-2})^{-1}$. Hence,

$$\begin{aligned} T_1^{(t+1)} &= \sum \left[w_i^{(t)} \mu^{(t)} + (1 - w_i^{(t)}) \bar{y}_i \right], \\ T_2^{(t+1)} &= \sum \left[w_i^{(t)} \mu^{(t)} + (1 - w_i^{(t)}) \bar{y}_i \right]^2 + \sum v_i^{(t)}, \\ T_3^{(t+1)} &= \sum_{i,j} (y_{ij} - \bar{y}_i)^2 + \sum_i n_i \left[w_i^{(t)} (\mu^{(t)} - \bar{y}_i)^2 + v_i^{(t)} \right]. \end{aligned} \quad (11.14)$$

The ML estimates based on complete data are

$$\begin{aligned} \hat{\mu} &= T_1/K, \\ \hat{\sigma}_\alpha^2 &= T_2/K - \hat{\mu}^2, \\ \hat{\sigma}_e^2 &= T_3/\sum_i n_i. \end{aligned} \quad (11.15)$$

Thus, the M step of EM replaces T_j by $T_j^{(t+1)}$ in these expressions, for $j = 1, \dots, 3$.

ML estimates from this algorithm are $\hat{\mu} = 53.3184$, $\hat{\sigma}_\alpha^2 = 54.8223$, and $\hat{\sigma}_e^2 = 249.2235$. The latter two estimates can be compared with $\tilde{\sigma}_\alpha^2 = 53.8740$, $\tilde{\sigma}_e^2 = 248.1876$, obtained by equating observed and expected mean squares from a random-effects ANOVA (e.g., see Brownlee 1965, section 11.4). Far more complex variance-components models can be fit using EM including those with multivariate y_{ij} , α_i , and X variables; see, e.g., Dempster et al. (1981) and Laird and Ware (1982). Gelfand et al. (1990) consider Bayesian inference for normal random effects models.

11.4 Multiple Linear Regression

11.4.1 Linear Regression with Missingness Confined to the Dependent Variable

Suppose a scalar outcome variable Y is regressed on p predictor variables X_1, \dots, X_p and missing values are confined to Y . If the missingness mechanism is ignorable, the incomplete observations do not contain information about the regression parameters, $\theta_{Y,X} = (\beta_{Y,X}, \sigma_{Y,X}^2)$. Nevertheless, the EM algorithm

can be applied to all observations and will obtain iteratively the same ML estimates as would have been obtained noniteratively using only the complete observations. Somewhat surprisingly, it may be easier to find these ML estimates iteratively by EM than noniteratively.

Example 11.5 Missing Outcomes in ANOVA. In designed experiments, the set of values of (X_1, \dots, X_p) is chosen to simplify the computation of least squares estimates. When Y given (X_1, \dots, X_p) is normal, least squares computations yield ML estimates. When values of Y , say y_i , $i = 1, \dots, m$, are missing, the remaining complete observations no longer have the balance occurring in the original design with the result that ML (least squares) estimation is more complicated. For a variety of reasons given in Chapter 2, it can be desirable to retain all observations and treat the problem as one with missing data.

When EM is applied to this problem, the M step corresponds to the least squares analysis on the original design and the E step involves finding the expected values and expected squared values of the missing y_i 's given the current estimated parameters $\theta_{Y \cdot X}^{(t)} = (\beta_{Y \cdot X}^{(t)}, \sigma_{Y \cdot X}^{(t)2})$:

$$y_i^{(t+1)} = E(y_i | X, Y_{\text{obs}}, \theta_{Y \cdot X}^{(t)}) = \begin{cases} y_i, & \text{if } y_i \text{ is observed } (i = m + 1, \dots, n), \\ \beta_{Y \cdot X}^{(t)} x_i^T, & \text{if } y_i \text{ is missing } (i = 1, \dots, m), \end{cases}$$

$$E(y_i^2 | X, Y_{\text{obs}}, \theta_{Y \cdot X}^{(t)}) = \begin{cases} y_i^2, & \text{if } y_i \text{ is observed,} \\ (\beta_{Y \cdot X}^{(t)} x_i^T)^2 + \sigma_{Y \cdot X}^{(t)2}, & \text{if } y_i \text{ is missing,} \end{cases}$$

where X is the $(n \times p)$ matrix of X values. Let Y be the $(n \times 1)$ vector of Y values, and $Y^{(t+1)}$ the vector Y with missing components y_i replaced by estimates from the E step at iteration $t + 1$. The M step calculates

$$\beta_{Y \cdot X}^{(t+1)} = (X^T X)^{-1} X^T Y^{(t+1)}, \quad (11.16)$$

and

$$\sigma_{Y \cdot X}^{(t+1)2} = n^{-1} \left[\sum_{i=m+1}^n (y_i - \beta_{Y \cdot X}^{(t)} x_i)^2 + m \sigma_{Y \cdot X}^{(t)2} \right]. \quad (11.17)$$

The algorithm can be simplified by noting that Eq. (11.16) does not involve $\sigma_{Y \cdot X}^{(t)2}$, and that at convergence we have

$$\sigma_{Y \cdot X}^{(t+1)2} = \sigma_{Y \cdot X}^{(t)2} = \hat{\sigma}_{Y \cdot X}^2,$$

so from (11.17)

$$\hat{\sigma}_{Y \cdot X}^2 = \frac{1}{n} \sum_{m+1}^n (y_i - \hat{\beta}_{Y \cdot X} x_i)^2 + \frac{m}{n} \hat{\sigma}_{Y \cdot X}^2,$$

or

$$\hat{\sigma}_{Y \cdot X}^2 = \frac{1}{n-m} \sum_{m+1}^n (y_i - \hat{\beta}_{Y \cdot X} x_i)^2. \quad (11.18)$$

Consequently, the EM iterations can omit the M step estimation of $\sigma_{Y \cdot X}^2$ and the E step estimation of $E(y_i^2 | \text{data}, \theta_{Y \cdot X}^{(t)})$, and find $\hat{\beta}_{Y \cdot X}$ by iteration. After convergence, we can calculate $\hat{\sigma}_{Y \cdot X}^2$ directly from (11.18). These iterations, which fill in the missing data, re-estimate the missing values from the ANOVA, and so forth, comprise the algorithm of Healy and Westmacott (1956) discussed in Section 2.4.3, with the additional correction for the degrees of freedom when estimating $\sigma_{Y \cdot X}^2$, obtained by replacing $n-m$ in Eq. (11.18) by $n-m-p$, to obtain the usual unbiased estimate of $\sigma_{Y \cdot X}^2$.

11.4.2 More General Linear Regression Problems with Missing Data

In general, there can be missing values in the predictor variables as well as in the outcome variable. For the moment, assume joint multivariate normality for (Y, X_1, \dots, X_p) . Then, applying Property 6.1, ML estimates or draws from the posterior distribution of the parameters of the regression of Y on X_1, \dots, X_p are standard functions of the ML estimates or posterior draws of the parameters of the multivariate normal distribution, discussed in the Section 11.2. Let

$$\theta = \begin{bmatrix} -1 & \mu_1 & \cdots & \mu_{p+1} \\ \mu_1 & \sigma_{11} & \cdots & \sigma_{1,p+1} \\ \vdots & \vdots & & \vdots \\ \mu_{p+1} & \sigma_{1,p+1} & & \sigma_{p+1,p+1} \end{bmatrix} \quad (11.19)$$

denote the augmented covariance matrix corresponding to the variables X_1, \dots, X_p and $X_{p+1} \equiv Y$. The intercept, slopes, and residual variance for the regression of Y on X_1, \dots, X_p are found in the last column of the matrix $\text{SWP}[1, \dots, p]\theta$, where the constant term and the predictor variables have been swept out of the matrix θ . Hence, if $\hat{\theta}$ is the ML estimate of θ found by the methods of Section 11.2, then ML estimates of the intercept, slopes, and residual variance are found from the last column of $\text{SWP}[1, \dots, p]\hat{\theta}$. Similarly, if $\theta^{(d)}$ is a draw from the posterior distribution of θ , then $\text{SWP}[1, \dots, p]\theta^{(d)}$ yields a draw from the joint posterior distribution of the intercept, slopes, and residual variance.

Let $\hat{\beta}_{YX \cdot X}$ and $\hat{\sigma}_{YY \cdot X}$ be the ML estimates of the regression coefficient of Y on X and residual variance of Y given X , respectively, as found by the EM algorithm just described. These estimates are ML under weaker conditions than multivariate normality of Y and (X_1, \dots, X_p) . Specifically, suppose we partition (X_1, \dots, X_p) as $(X_{(A)}, X_{(B)})$, where the variables in $X_{(A)}$ are more observed than both Y and the variables in $X_{(B)}$ in the sense of Section 7.5 that any unit with any observation on Y or $X_{(B)}$ has all variables in $X_{(A)}$ observed. A particularly simple case occurs when $X = (X_1, \dots, X_p)$ is fully observed so that $X_{(A)} = X$: see Figure 7.1 for the general case, where Y_1 corresponds to $(Y, X_{(B)})$ and Y_3 corresponds to $X_{(A)}$ and Y_2 is null. Then $\hat{\beta}_{YX \cdot X}$ and $\hat{\sigma}_{YY \cdot X}$ are also ML estimates if the conditional distribution of $(Y, X_{(B)})$ given $X_{(A)}$ is multivariate normal – see Chapter 7 for details. This conditional multivariate normal assumption is much less stringent than multivariate normality for X_1, \dots, X_{p+1} , because it allows the predictors in $X_{(A)}$ to be categorical variables, as in “dummy variable regression,” and also allows interactions and polynomials in the completely observed predictors to be introduced into the regression without affecting the propriety of the incomplete data procedure. Similarly for Bayesian inference, if $\beta_{YX \cdot X}^{(d)}, \sigma_{YX \cdot X}^{(d)}$ are draws from the posterior distribution from multivariate normal DA algorithm, they are also draws from the posterior distribution under a conditional multivariate normal model for $(Y, X_{(B)})$ given $X_{(A)}$.

The $(p \times p)$ submatrix of the first p rows and columns of $\text{SWP}[1, \dots, p]\hat{\theta}$ does not provide the asymptotic covariance matrix of the estimated regression coefficients, as is the case with complete data. The asymptotic covariance matrix of the estimated slopes based on the usual large-sample approximation generally involves the inversion of the full information matrix of the means, variances, and covariances, which is displayed in Section 11.2.2. Computationally, simpler alternatives are to apply the bootstrap, or to simulate the posterior distribution of the parameters. In particular, the set of draws $\text{SWP}[1, \dots, p]\theta^{(d)}$, where $\theta^{(d)}$ is a draw from the posterior distribution of θ , can be used to simulate the posterior distribution of $\text{SWP}[1, \dots, p]\theta$, thereby allowing the construction of posterior credibility intervals for the regression coefficients and residual variance.

More generally, ML or Bayes estimation for multivariate linear regression can be achieved by applying the algorithms of Section 11.2, and then sweeping the independent variables in the resulting augmented covariance matrix. Specifically, if the dependent variables are Y_1, \dots, Y_K and the independent variables are X_1, \dots, X_p , then the augmented covariance matrix of the combined set of variables $(X_1, \dots, X_p, Y_1, \dots, Y_K)$ is estimated using the multivariate normal EM algorithm, and then the variables X_1, \dots, X_p are swept in the matrix. The resulting matrix contains the ML estimates of the $(p \times K)$ matrix of regression coefficients of Y on X and the $(K \times K)$ residual covariance matrix of Y given X . The parallel operations on draws from the posterior distribution by DA provide

draws of the multivariate regression parameters. For a review of these methods and alternatives, see Little (1992).

Example 11.6 *MANOVA with Missing Data Illustrated Using the St. Louis Data (Example 11.1 Continued).* We now apply the multivariate normal model to all the data in Table 11.1, including an indicator variable for the low and medium/high-risk groups, and then sweep the group indicator variable out of the augmented covariance matrix at the final step to yield estimates from the multivariate regression of the continuous outcomes on the group indicator variables. The regression coefficient of the group indicator measures the difference in mean outcome between the low-, medium- and high-risk groups. Figure 11.2 displays histograms of 9000 draws of these regression coefficients from DA. The 95% posterior probability intervals based on the 2.5th–97.5th

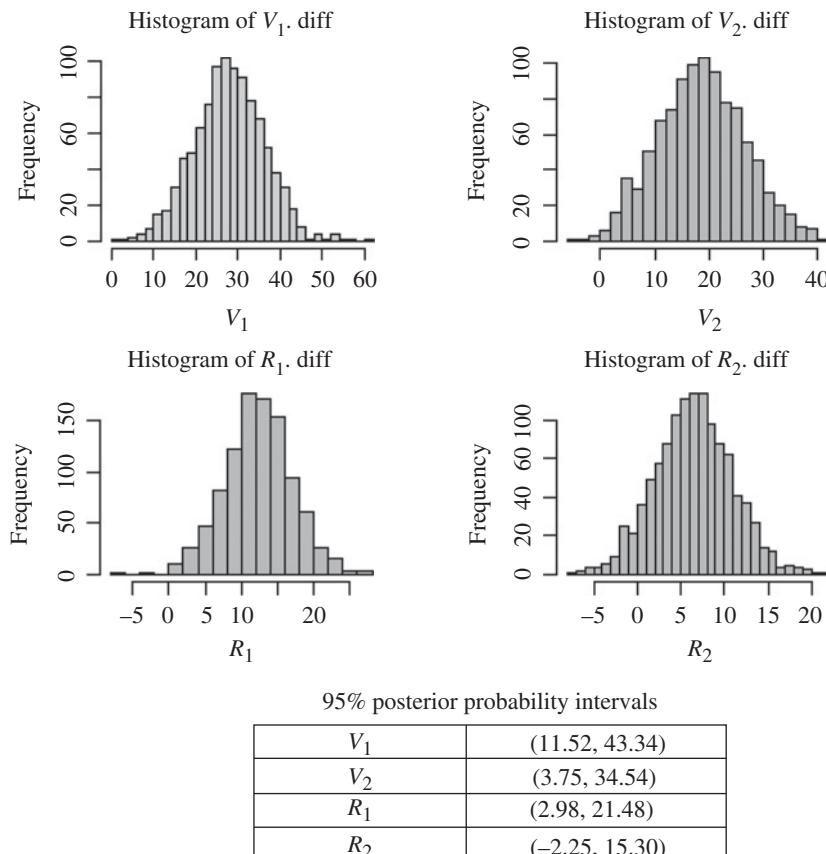


Figure 11.2 Example 11.6, posterior distributions of mean differences $\mu_{\text{low}} - \mu_{\text{med/high}}$, St. Louis risk research data, based on 9000 draws, multivariate normal regression model.

percentiles are shown in the plots in Figure 11.2. Conclusions are similar to Example 11.1, namely, reading and verbal means appear higher in the low-risk group than in the moderate- and high-risk group.

11.5 A General Repeated-Measures Model with Missing Data

Missing data often occur in longitudinal studies, where subjects are observed at different times and/or under different experimental conditions. Normal models for such data often combine special covariance structures such as those discussed in Section 11.3 with mean structures that relate the mean of the repeated measures to design variables. The following general repeated measures model is given in Jennrich and Schluchter (1986) and builds on earlier work by Harville (1977), Laird and Ware (1982), and Ware (1985). ML for this model has been implemented in a number of software programs, including SAS (1992) and S-Plus (Schafer 1998; Pinheiro and Bates 2000).

Suppose that the hypothetical complete data for unit i consist of K measurements $y_i = (y_{i1}, \dots, y_{iK})$ on an outcome variable Y , and

$$y_i \sim_{\text{ind}} N_K(X_i\beta, \Sigma(\psi)), \quad (11.20)$$

where X_i is a known $(K \times m)$ design matrix for unit i , β is a $(m \times 1)$ vector of unknown regression coefficients, and the elements of the covariance matrix Σ are known functions of a set of v unknown parameters ψ . The model thus incorporates a mean structure, defined by the set of design matrices $\{X_i\}$, and a covariance structure, defined by the form of the covariance matrix Σ . The observed data consist of the design matrices $\{X_i\}$ and $\{y_{(0),i} : i = 1, \dots, n\}$ where $y_{(0),i}$ is the observed part of the vector y_i . Missing values of y_i are assumed to be ignorable. The complete-data loglikelihood is linear in the quantities $\{y_i, y_i^T y_i : i = 1, \dots, n\}$. Hence, the E step consists in calculating the means of y_i and $y_i^T y_i$ given $y_{(0),i}$, X_i , and current estimates of β and Σ . These calculations involve sweep operations on the current estimate of Σ analogous to those in the multivariate normal model of Section 11.2.1. The M step for the model is itself iterative except in special cases, and thus a primary attraction of EM, the simplicity of the M step, is lost. Jennrich and Schluchter (1986) present a generalized EM algorithm (see Section 8.4) and also discuss scoring and Newton–Raphson algorithms that can be attractive when Σ depends on a modest number of parameters, ψ .

A large number of situations can be modeled by combining different choices of mean and covariance structures, for example:

Independence: $\Sigma = \text{Diag}_K(\psi_1, \dots, \psi_K)$, a diagonal $(K \times K)$ matrix with entries $\{\psi_k\}$,

Compound symmetry: $\Sigma = \psi_1 U_K + \psi_2 I_K$, ψ_1 and ψ_2 scalar, $U_K = (K \times K)$ matrix of ones, $I_K = (K \times K)$ identity matrix,

Autoregressive lag 1 (AR1): $\Sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_1 \psi_2^{|j-k|}$, ψ_1 , ψ_2 scalars,

Banded: $\Sigma = (\sigma_{jk})$, $\sigma_{jk} = \psi_a$, where $a = |j - k| + 1$, $a = 1, \dots, K$,

Factor analytic: $\Sigma = \Gamma \Gamma^T + \psi$, $\Gamma = (K \times q)$ matrix of unknown factor loadings, and $\psi = (K \times K)$ diagonal matrix of “specific variances,”

Random effects: $\Sigma = Z\psi Z^T + \sigma^2 I_K$, $Z = (K \times q)$ known matrix, $\psi = (q \times q)$ unknown dispersion matrix, σ^2 scalar, I_K the $K \times K$ identity matrix,

Unstructured: $\Sigma = (\sigma_{jk})$, $\psi_1 = \sigma_{11}, \dots, \psi_K = \sigma_{1K}$, $\psi_{K+1} = \sigma_{22}, \dots, \psi_v = \sigma_{KK}$, $v = K(K+1)/2$.

The mean structure is also very flexible. If $X_i = I_K$, then $\mu_i = \beta^T$ for all i . This constant mean structure, combined with the unstructured, factor analytic and compound symmetry covariance structures, yields the models of Section 11.2, Examples 11.3 and 11.4, respectively. Between-subject and within-subject effects are readily modeled through other choices of X_i , as in the next example.

Example 11.7 Growth-Curve Models with Missing Data. Potthoff and Roy (1964) present the growth data in Table 11.4 for 11 girls and 16 boys. For each subject, the distance from the center of the pituitary to the maxillary fissure was recorded at the ages of 8, 10, 12, and 14. Jennrich and Schluchter (1986) fit eight repeated-measures models to these data. We fit the same models to the data obtained by deleting the 10 values in parentheses in Table 11.4. The missingness mechanism is designed to be MAR, but not MCAR. Specifically, for both girls and boys, values at age 10 are deleted for cases with low values at age 8. Table 11.5 summarizes the models, giving values of minus twice the loglikelihood (-2λ) and the likelihood ratio chi-squared (χ^2) for comparing models. The last column gives values for the latter statistic from the complete data before deletion, as given in Jennrich and Schluchter (1986).

For the i th subject, let y_i denote the four distance measurements, and let x_i be a design variable equal to 1 if the child is a boy and 0 if the child is a girl. Model 1 specifies a distinct mean for each of the sex by age groups, and assumes that the (4×4) covariance matrix is unstructured. The X_i matrix for subject i can be written as

$$X_i = \begin{bmatrix} 1 & x_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & x_i & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & x_i \end{bmatrix}.$$

Table 11.4 Example 11.7, growth data for 11 girls and 16 boys

Individual girl	Age (in years)				Individual boy	Age (in years)			
	8	10	12	14		8	10	12	14
1	21	20	21.5	23	1	26	25	29	31
2	21	21.5	24	25.5	2	21.5	(22.5)	23	26.5
3	20.5	(24)	24.5	26	3	23	22.5	24	27.5
4	23.5	24.5	25	26.5	4	25.5	27.5	26.5	27
5	21.5	23	22.5	23.5	5	20	(23.5)	22.5	26
6	20	(21)	21	22.5	6	24.5	25.5	27	28.5
7	21.5	22.5	23	25	7	22	22	24.5	26.5
8	23	23	23.5	24	8	24	21.5	24.5	25.5
9	20	(21)	22	21.5	9	23	20.5	31	26.0
10	16.5	(19)	19	19.5	10	27.5	28	31	31.5
11	24.5	25	28	28	11	23	23	23.5	25
					12	21.5	(23.5)	24	28
					13	17	(24.5)	26	29.5
					14	22.5	25.5	25.5	26
					15	23	24.5	26	30
					16	22	(21.5)	23.5	(25)

Values in parentheses are treated as missing in Example 11.6.

Source: Potthoff and Roy (1964) as reported by Jennrich and Schluchter (1986). Reproduced with permission of Oxford University Press.

With no missing data, the ML estimate of β is the vector of eight sample means and the ML estimate of Σ is S/n , where S is the pooled within-groups sum of squares and cross-products matrix.

This unrestricted model, Model 1 in Table 11.5, was fitted to the incomplete data of Table 11.4. Seven other models were also fitted to those data. Plots suggest a linear relationship between mean distance and age, with different intercepts and slopes for girls and boys. The mean structure for this model can be written as

$$\mu_i^T = X_i \beta = \begin{bmatrix} 1 & x_i & -3 & -3x_i \\ 1 & x_i & -1 & -x_i \\ 1 & x_i & 1 & x_i \\ 1 & x_i & 3 & 3x_i \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \quad (11.21)$$

Table 11.5 Example 11.7, summary of models fitted

Model number	Description	Number of parameters	-2λ	Comparison model	χ^2	df	Complete data χ^2
1	Eight separate means, unstructured covariance matrix	18	386.96	—	—	—	—
2	Two lines, unequal slopes, unstructured covariance matrix	14	393.29	1	6.33	4	[2.97]
3	Two lines, common slope, unstructured covariance matrix	13	397.40	2	4.11	1	[6.68]
4	Two lines, unequal slopes, banded structure	8	398.03	2	4.74	6	[5.17]
5	Two lines, unequal slopes, AR(1) structure	6	409.52	2	16.24	8	[21.20]
6	Two lines, unequal slopes, random slopes and intercepts	8	400.45	2	7.16	6	[8.33]
7	Two lines, unequal slopes, random intercepts (compound symmetry)	6	401.31	2	8.02	8	[9.16]
8	Two lines, unequal slopes, independent observations	5	441.58	7	40.27	1	[50.83]

Source: The complete data is obtained from Jennrich and Schluchter (1986). Reproduced with permission of John Wiley and Sons.

where β_1 and $\beta_1 + \beta_2$ represent overall means and β_3 and $\beta_3 + \beta_4$ represent slopes for girls and boys, respectively. Model 2 fits this mean structure with an unstructured Σ .

The likelihood-ratio statistic comparing Model 2 with Model 1 is $\chi^2 = 6.33$ on 4 degrees of freedom, indicating a fairly satisfactory fit for Model 2 relative to Model 1. Model 3 is obtained from Model 2 by setting $\beta_4 = 0$, that is, dropping the last column of X_i . It constrains the regression lines of distance against age to have common slope in the two groups. Compared with Model 2, Model 3 yields a likelihood ratio of 4.11 on 1 degree of freedom, indicating significant lack of fit. Hence, the mean structure of Model 2 is preferred.

The remaining models in Table 11.5 have the mean structure of Model 2, but place constraints on Σ . The autoregressive (Model 5) and independence

(Model 8) covariance structures do not fit the data, judging from the chi-squared statistics. The banded structure (Model 4) and two random effects structures (Models 6 and 7) fit the data well. Of these, Model 7 may be preferred on grounds of parsimony. The model can be interpreted as a random effects model with a fixed slope for each sex group and a random intercept that varies across subjects about common means for boys and girls. Further analysis would display the parameter estimates for this preferred model.

11.6 Time Series Models

11.6.1 Introduction

We confine our limited discussion of time-series modeling with missing data to parametric time-domain models with normal disturbances, because these models are most amenable to the ML techniques developed in Chapters 6 and 8. Two classes of models of this type appear particularly important in applications: the autoregressive-moving average (ARMA) models developed by Box and Jenkins (1976), and general state-space or Kalman-filter models, initiated in the engineering literature (Kalman 1960) and enjoying considerable development in the econometrics and statistics literature on time series (Harvey 1981). As discussed in the next section, autoregressive models are relatively easy to fit to incomplete time-series data, with the aid of the EM algorithm. Box–Jenkins models with moving average components are less easily handled, but ML estimation can be achieved by recasting the models as general state-space models, as discussed in Harvey and Phillips (1979) and Jones (1980). The details of this transformation are omitted here; however, ML estimation for general state-space models from incomplete data is outlined in Section 11.6.3, following the approach of Shumway and Stoffer (1982).

11.6.2 Autoregressive Models for Univariate Time Series with Missing Values

Let $Y = (y_0, y_1, \dots, y_T)$ denote a completely observed univariate time series with $T + 1$ observations. The autoregressive model of lag p (AR p) assumes that y_i , the value at time i , is related to values at p previous time points by the model

$$(y_i | y_1, y_2, \dots, y_{i-1}, \theta) \sim N(\alpha + \beta_1 y_{i-1} + \dots + \beta_p y_{i-p}, \sigma^2), \quad (11.22)$$

where $\theta = (\alpha, \beta_1, \beta_2, \dots, \beta_p, \sigma^2)$, α is a constant term, $\beta_1, \beta_2, \dots, \beta_p$ are unknown regression coefficients, and σ^2 is an unknown error variance. Least squares estimates of $\alpha, \beta_1, \beta_2, \dots, \beta_p$ and σ^2 can be found by regressing y_i on $x_i = (y_{i-1}, y_{i-2}, \dots, y_{i-p})$, using observations $i = p, p+1, \dots, T$. These estimates are only approximately ML because the contribution of the marginal distribution of

y_0, y_1, \dots, y_{p-1} to the likelihood is ignored, which is justified when p is small compared with T .

If some observations in the series are missing, one might consider applying the methods of Section 11.4 for regression with missing values. This approach may yield useful rough approximations, but the procedure is not ML, even assuming the marginal distribution of y_0, y_1, \dots, y_{p-1} can be ignored, because (i) missing values y_i ($i \geq p$) appear as dependent and independent variables in the regressions, and (ii) the model (11.22) induces a special structure on the mean vector and covariance matrix of Y that is not used in the analysis. Thus, special EM algorithms are required to estimate the AR p model from incomplete time series. The algorithms are relatively easy to implement, although not trivial to describe. We confine attention here to the $p = 1$ case.

Example 11.8 *The Autoregressive Lag 1 (AR1) Model for Time Series with Missing Values.* Setting $p = 1$ in Eq. (11.22), we obtain the model

$$(y_i | y_1, \dots, y_{i-1}, \theta) \sim_{\text{ind}} N(\alpha + \beta y_{i-1}, \sigma^2). \quad (11.23)$$

The AR1 series is *stationary*, yielding a constant marginal distribution of y_i over time, only if $|\beta| < 1$. The joint distribution of the y_i then has constant marginal mean $\mu \equiv \alpha(1 - \beta)^{-1}$, variance $\text{Var}(y_i) = \sigma^2(1 - \beta^2)^{-1}$, and covariances $\text{Cov}(y_i, y_{i+k}) = \beta^k \sigma^2(1 - \beta^2)^{-1}$ for $k \geq 1$. Ignoring the contribution of the marginal distribution of y_0 , the complete-data loglikelihood for Y is $\ell(\alpha, \beta, \sigma^2 | y) = -0.5\sigma^{-2} \sum_{i=1}^T (y_i - \alpha - \beta y_{i-1})^2 - 0.5T \ln \sigma^2$, which is equivalent to the log-likelihood for the normal linear regression of y_i on $x_i = y_{i-1}$, with data $\{(y_i, x_i), i = 1, \dots, T\}$. The complete-data sufficient statistics are $s = (s_1, s_2, s_3, s_4, s_5)$, where

$$s_1 = \sum_{i=1}^T y_i, \quad s_2 = \sum_{i=1}^T y_{i-1}, \quad s_3 = \sum_{i=1}^T y_i^2, \quad s_4 = \sum_{i=1}^T y_{i-1}^2, \quad s_5 = \sum_{i=1}^T y_i y_{i-1}.$$

ML estimates of $\theta = (\alpha, \beta, \sigma)$ are $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$, where

$$\begin{aligned} \hat{\alpha} &= (s_1 - \hat{\beta}s_2)T^{-1}, \\ \hat{\beta} &= (s_5 - T^{-1}s_1s_2)(s_4 - T^{-1}s_2^2)^{-1}, \\ \hat{\sigma}^2 &= \{s_3 - s_1^2 T^{-1} - \hat{\beta}^2(s_4 - s_2^2 T^{-1})\} / T. \end{aligned} \quad (11.24)$$

Now suppose some observations are missing, and missingness is ignorable. ML estimates of θ , still ignoring the contribution of the marginal distribution of y_0 to the likelihood, can be obtained by the EM algorithm. Let $\theta^{(t)} = (\alpha^{(t)}, \beta^{(t)}, \sigma^{(t)})$ be estimates of θ at iteration t . The M step of the algorithm calculates $\theta^{(t+1)}$ from Eq. (11.24) with complete data sufficient statistics s replaced by estimates $s^{(t)}$ from the E step.

The E step computes $s^{(t)} = (s_1^{(t)}, s_2^{(t)}, s_3^{(t)}, s_4^{(t)}, s_5^{(t)})$, where

$$\begin{aligned}s_1^{(t)} &= \sum_{i=1}^T \hat{y}_i^{(t)}, \quad s_2^{(t)} = \sum_{i=1}^T \hat{y}_{i-1}^{(t)}, \quad s_3^{(t)} = \sum_{i=1}^T \left\{ \left(\hat{y}_i^{(t)} \right)^2 + c_{ii}^{(t)} \right\}, \\ s_4^{(t)} &= \sum_{i=1}^T \left\{ \left(\hat{y}_{i-1}^{(t)} \right)^2 + c_{i-1,i-1}^{(t)} \right\}, \quad s_5^{(t)} = \sum_{i=1}^T \left\{ \hat{y}_{i-1}^{(t)} \hat{y}_i^{(t)} + c_{i-1,i}^{(t)} \right\},\end{aligned}$$

and

$$\begin{aligned}\hat{y}_i^{(t)} &= \begin{cases} y_i, & \text{if } y_i \text{ is observed,} \\ E\{y_i \mid Y_{(0)}, \theta^{(t)}\}, & \text{if } y_i \text{ is missing,} \end{cases} \\ c_{ij}^{(t)} &= \begin{cases} 0, & \text{if } y_i \text{ or } y_j \text{ is observed,} \\ \text{Cov}\{y_i, y_j \mid Y_{(0)}, \theta^{(t)}\}, & \text{if } y_i \text{ and } y_j \text{ are missing.} \end{cases}\end{aligned}$$

The E step involves standard sweep operations on the covariance matrix of the observations. However, this $(T \times T)$ matrix is usually large, so it is desirable to exploit properties of the AR1 model to simplify the E step computations. Suppose $Y_{(1)}^* = (y_{j+1}, y_{j+2}, \dots, y_{k-1})$ is a sequence of missing values between observed values y_j and y_k . Then (i) $Y_{(1)}^*$ is independent of the other missing values, given $Y_{(0)}$ and θ , and (ii) the distribution of $Y_{(1)}^*$ given $Y_{(0)}$ and θ depends on $Y_{(0)}$ only through the bounding observations y_j and y_k . The latter distribution is multivariate normal, with constant covariance matrix, and means that are weighted averages of $\mu = \alpha(1 - \beta)^{-1}$, y_j and y_k . The weights and covariance matrix depend only on the number of missing values in the sequence and can be found from the current estimate of the covariance matrix of $(y_j, y_{j+1}, \dots, y_k)$ by sweeping on elements corresponding to the observed variables y_j and y_k .

In particular, suppose y_j and y_{j+2} are present and y_{j+1} is missing. The covariance matrix of y_j , y_{j+1} and y_{j+2} is

$$A = \frac{\sigma^2}{1 - \beta^2} \begin{bmatrix} 1 & \beta & \beta^2 \\ \beta & 1 & \beta \\ \beta^2 & \beta & 1 \end{bmatrix}.$$

Sweeping on y_j and y_{j+2} yields

$$\text{SWP}[j, j+2]A = \frac{1}{1 + \beta^2} \begin{bmatrix} -\sigma^{-2} & \beta & -\beta^2\sigma^{-2} \\ \beta & \sigma^2 & \beta \\ -\beta^2\sigma^{-2} & \beta & -\sigma^{-2} \end{bmatrix}. \quad (11.25)$$

Hence, from stationarity and (11.25),

$$\begin{aligned} E\{y_{j+1} \mid y_j, y_{j+2}, \theta\} &= \mu + \beta(1 + \beta^2)^{-1}(y_{j+2} - \mu) + \beta(1 + \beta^2)^{-1}(y_j - \mu) \\ &= \mu \left\{ 1 - \frac{2\beta}{1 + \beta^2} \right\} + \frac{\beta}{1 + \beta^2}\{y_j + y_{j+2}\}, \\ \text{Var}(y_{j+1} \mid y_j, y_{j+2}, \theta) &= \sigma^2(1 + \beta^2)^{-1}. \end{aligned}$$

Substituting $\theta = \theta^{(t)}$ in these expressions yields $\hat{y}_{j+1}^{(t)}$ and $\hat{c}_{j+1,j+1}^{(t)}$ for the E step.

11.6.3 Kalman Filter Models

Shumway and Stoffer (1982) consider the Kalman filter model

$$\begin{aligned} (y_i \mid A_i, z_i, \theta) &\sim_{\text{ind}} N(z_i A_i, B), \\ (z_0 \mid \theta) &\sim N(\mu, \Sigma), \\ (z_i \mid z_1, \dots, z_{i-1}, \theta) &\sim N(z_{i-1} \phi, Q), i \geq 1, \end{aligned} \tag{11.26}$$

where y_i is a $(1 \times q)$ vector of observed variables at time i , A_i is a known $(p \times q)$ matrix that relates the mean of y_i to an unobserved $(1 \times p)$ stochastic vector z_i , and $\theta = (B, \mu, \Sigma, \phi, Q)$ represents the unknown parameters, where B , Σ , and Q are covariance matrices, μ is the mean of z_0 , and ϕ is a $(p \times p)$ matrix of autoregression coefficients of z_i on z_{i-1} . The random unobserved series z_i , which is modeled as a first-order multivariate autoregressive process, is of primary interest.

This model can be envisioned as a kind of random effects model for time series, where the effect vector z_i has correlation structure over time. The primary aim is to predict the unobserved series $\{z_i\}$ for $i = 1, 2, \dots, n$ (smoothing) and for $i = n+1, n+2, \dots$ (forecasting), using the observed series y_1, y_2, \dots, y_n . If the parameter θ were known, the standard estimates of z_i would be their conditional means, given the parameters θ and the data Y . These quantities are called Kalman smoothing estimators, and the set of recursive formulas used to derive them are called the Kalman filter. In practice, θ is unknown, and the forecasting and smoothing procedures involve ML estimation of θ , and then application of the Kalman filter with θ replaced by the ML estimate $\hat{\theta}$.

The same process applies when data Y are incomplete, with Y replaced by its observed component, say $Y_{(0)}$. ML estimates of Q can be derived by Newton-Raphson techniques (Gupta and Mehra 1974; Ledolter 1979; Goodrich and Caines 1979). However, the EM algorithm provides a convenient alternative method, with the missing components $Y_{(1)}$ of Y and z_1, z_2, \dots, z_n treated as missing data. An attractive feature of this approach is that the E step of the algorithm includes the calculation of the expected value of z_i given $Y_{(0)}$ and current

estimates of θ , which is the same process as Kalman smoothing described above. Details of the E step are given in Shumway and Stoffer (1982). The M step is relatively straightforward. Estimates of ϕ and Q are obtained by autoregression applied to the expected values of the complete data sufficient statistics

$$\sum_{i=1}^n z_i, \quad \sum_{i=1}^n z_i^T z_i, \quad \sum_{i=1}^n z_{i-1}, \quad \sum_{i=1}^n z_{i-1}^T z_{i-1}, \quad \text{and} \quad \sum_{i=1}^n z_{i-1}^T z_i$$

from the E step; B is estimated by the expected value of the residual covariance matrix $n^{-1} \sum_{i=1}^n (y_i - z_i A_i)^T (y_i - z_i A_i)$. Finally, μ is estimated as the expected value of z_0 , and Σ is set from external considerations. We now provide a specific example of this very general model.

Example 11.9 *A Bivariate Time Series Measuring an Underlying Series with Error.* Table 11.6, taken from Meltzer et al. (1980), shows two incomplete time series of total expenditures for physician services, measured by the Social Security Administration (SSA), yielding Y_1 , and the Health Care Financing Administration (HCFA), yielding Y_2 . Shumway and Stoffer (1982) analyze the data using the model

$$(y_{ij} | z_i, \theta) \sim_{\text{ind}} N(z_i, B_j), \quad i = 1949, \dots, 1981, \\ (z_i | z_1, \dots, z_{i-1}, \theta) \sim N(z_{i-1}\phi, Q), \quad i = 1950, \dots, 1981,$$

where y_{ij} is the total expenditure amount at time i for SSA ($j = 1$) and HCFA ($j = 2$), z_i is the underlying true expenditure, assumed to form an AR1 series over time with coefficient ϕ and residual variance Q , B_j is the measurement variance of y_{ij} ($j = 1, 2$), and $\theta = (B_1, B_2, \phi, Q)$. Unlike Example 11.8, the AR1 series for z_i is not assumed stationary, the parameter ϕ being an inflation factor modeling exponential growth; the assumption that ϕ is constant over time is probably an oversimplification. The last columns of Table 11.6 show smoothed estimates of z_i from the final iteration of the EM algorithm for years 1949–1976, and predictions for the five years 1977–1981, together with their standard errors. The predictions for 1977–1981 have standard errors ranging from 355 for 1977 to 952 for 1982, reflecting considerable uncertainty.

11.7 Measurement Error Formulated as Missing Data

In Chapter 1, we described how measurement error can be formulated as a missing-data problem, where the true values of a variable measured with error are treated as completely missing. Guo and Little (2011) apply this idea to internal calibration data with heteroskedastic measurement error. In our final

Table 11.6 Example 11.9, data set and predictions from the EM algorithm – physician expenditures (in millions)

Year (i)	SSA		Predictions from EM algorithm	
	y_{i1}	y_{i2}	$E(z_i Y_{(0)}, \theta)$	$\text{Var}^{1/2}(z_i Y_{(0)}, \theta)$
1949	2 633	—	2 541	178
1950	2 747	—	2 711	185
1951	2 868	—	2 864	186
1952	3 042	—	3 045	186
1953	3 278	—	3 269	186
1954	3 574	—	3 519	186
1955	3 689	—	3 736	186
1956	4 067	—	4 063	186
1957	4 419	—	4 433	186
1958	4 910	—	4 876	186
1959	5 481	—	5 331	186
1960	5 684	—	5 644	186
1961	5 895	—	5 972	186
1962	6 498	—	6 477	186
1963	6 891	—	7 032	185
1964	8 065	—	7 866	179
1965	8 745	8 474	8 521	110
1966	9 156	9 175	9 198	108
1967	10 287	10 142	10 160	108
1968	11 099	11 104	11 159	108
1969	12 629	12 648	12 645	108
1970	14 306	14 340	14 289	108
1971	15 835	15 918	15 835	108
1972	16 916	17 162	17 171	108
1973	18 200	19 278	19 106	109
1974	—	21 568	21 675	119
1975	—	25 181	25 027	120
1976	—	27 931	27 932	129
1977	—		31 178	355
1978		—	34 801	512
1979		—	38 846	657
1980		—	43 361	802
1981		—	48 400	952

Source: Meltzer et al. (1980) as reported in Shumway and Stoffer (1982), Tables I and III.
Reproduced with permission of John Wiley and Sons.

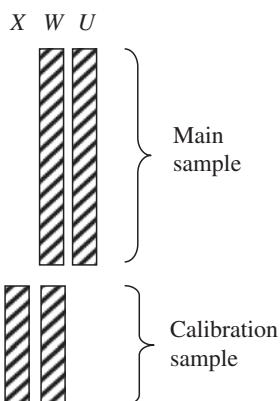


Figure 11.3 Missingness pattern for Example 11.10, with shaded values observed. X = true covariate, missing in the main sample; W = measured covariate, observed in both the main and calibration samples; U = other variables, missing in the calibration sample.

example in this chapter, we describe MI to address measurement error for data from a main sample and an external calibration sample, described in Example 1.15. For more details, see Guo et al. (2011).

Example 11.10 Measurement Error as Missing Data: A Normal Model for External Calibration. In Example 1.15, we described data displayed in Figure 11.3, where the main sample data are a random sample of values of U and W , where W is the proxy variable for X , and information relating W and X is obtained from an external calibration sample in which values of X and W are recorded. Here X and W are univariate, U is a vector of p variables, and interest concerns parameters of the joint distribution of X and U . The missingness pattern is similar to that of the file-matching problem described in Example 1.7. An important special case is where $U = (Y, Z)$, where Y is a vector of q outcomes, Z is a vector of r covariates, $p = q + r$, and interest lies in the regression of Y on Z and X . This pattern arises in the case of external calibration, where calibration of W is carried out independently of the main study, for example by an assay conducted by the manufacturer. Typically, data from the calibration sample are not available to the analyst, but we assume that summary statistics – namely the mean and covariance matrix of X and W – are available. We assume the missing data, namely the values of X in the main sample, and the missing values of Y and Z in the calibration sample, are ignorable.

Guo et al. (2011) assume that in the main sample and the calibration sample, the conditional distribution of U and X given W is $(p + 1)$ -variate normal with a mean that is linear in W and a constant covariance matrix. Further, this conditional distribution is assumed to be the same in the main study sample and the calibration sample, although the distribution of W can differ in the two

samples. This indispensable assumption is a form of the “transportability across studies” assumption in Carroll et al. (2006). It is evident from Figure 11.3 that the joint distribution cannot be estimated from the data without invoking more assumptions, because the variables X and U are never jointly observed. Specifically, there is no information about the p partial correlations of X and U given Z .

To address this issue, we make the “nondifferential measurement error” (NDME) assumption, which states that the distribution of U given W and X does not depend on W .

This assumption is reasonable if the measurement error in W is unrelated to values of $U = (Y, Z)$, and is plausible in some bioassays. Our approach multiply-imputes the missing values of X in the main study from the estimated conditional distribution of X given the observed variables in the main study sample, namely U and W ; under normality assumptions, this is surprisingly straightforward, computationally. Let $\theta_{x\cdot uw} = (\beta_{x\cdot uw}, \sigma_{x\cdot uw})$ where $\beta_{x\cdot uw}$ denotes the vector of regression coefficients and $\sigma_{x\cdot uw}$ the residual standard deviation for that regression. For data set d , a draw $\theta_{x\cdot uw}^{(d)} = (\beta_{x\cdot uw}^{(d)}, \sigma_{x\cdot uw}^{(d)})$ is taken from the posterior distribution of $\theta_{x\cdot uw}$. This draw can be computed rather simply from the main sample data and summary statistics from the external calibration sample, namely the sample size, sample mean and sum of squares and cross products matrix of X and W .

Specifically, let $\theta = (\beta_{xw\cdot w}, \sigma_{xx\cdot w}, \beta_{uw\cdot w}, \sigma_{uu\cdot w}, \sigma_{ux\cdot w})$, where $(\beta_{xw\cdot w}, \sigma_{xx\cdot w})$ are the regression coefficients and residual variance for the (normal) regression of X on W , $(\beta_{uw\cdot w}, \sigma_{uu\cdot w})$ are the regression coefficients and residual covariance matrix for the (normal) regression of U on W , and $\sigma_{ux\cdot w}$ represents the set of p partial covariances of U and X given W . Now:

- Draw $(\beta_{uw\cdot w}^{(d)}, \sigma_{uu\cdot w}^{(d)})$ from the distribution of $(\beta_{uw\cdot w}, \sigma_{uu\cdot w})$, given the data on U and W in the main study sample; and
- Draw $(\beta_{xw\cdot w}^{(d)}, \sigma_{xx\cdot w}^{(d)})$ from the distribution of $(\beta_{xw\cdot w}, \sigma_{xx\cdot w})$, given the data on X and W in the calibration sample. Note that these draws $(\beta_{xw\cdot w}^{(d)}, \sigma_{xx\cdot w}^{(d)})$ can be computed from summary statistics in the calibration sample, namely the sample size, sample mean, and sum of squares and cross-products matrix of X and W .

Both (a) and (b) are straightforward, because both these distributions are posterior distributions for complete-data problems, as discussed in Example 6.17. To obtain a draw for the remaining component of θ , namely $\sigma_{ux\cdot w}$, note that by properties of the normal distribution, the regression coefficient of W in the multivariate regression of U on X and W can be expressed as

$$\beta_{uw\cdot xw} = \beta_{uw\cdot w} - \sigma_{ux\cdot w}\beta_{xw\cdot w}/\sigma_{xx\cdot w}.$$

The NDME assumption implies that $\beta_{uw \cdot xw} = 0$. Hence

$$\beta_{uw \cdot w} - \sigma_{ux \cdot w} \beta_{xw \cdot w} / \sigma_{xx \cdot w} = 0, \text{ so } \sigma_{ux \cdot w} = \beta_{uw \cdot w} \sigma_{xx \cdot w} / \beta_{xw \cdot w}.$$

Thus we have expressed $\sigma_{ux \cdot w}$ as a function of the other parameters, and a draw of $\sigma_{ux \cdot w}$ is:

$$\sigma_{ux \cdot w}^{(d)} = \beta_{uw \cdot w}^{(d)} \sigma_{xx \cdot w}^{(d)} / \beta_{xw \cdot w}^{(d)}. \quad (11.27)$$

Combining, we thus have a draw $\theta^{(d)} = (\beta_{xw \cdot w}^{(d)}, \sigma_{xx \cdot w}^{(d)}, \beta_{uw \cdot w}^{(d)}, \sigma_{uu \cdot w}^{(d)}, \sigma_{ux \cdot w}^{(d)})$ from the conditional distribution of X and U given W . Missing values x_i of X for the i th observation in the study sample are then imputed as draws from the conditional normal distribution of X given U and W , with parameters $\beta_{x \cdot uw}^{(d)}$ and $\sigma_{xx \cdot uw}^{(d)}$, functions of $\theta^{(d)}$, obtained by sweeping out U to convert U from dependent to independent variables. That is

$$x_i^{(d)} = E(x_i | y_i, z_i, w_i, \beta_{x \cdot uw}^{(d)}) + e_i^{(d)} \sqrt{\sigma_{xx \cdot uw}^{(d)}},$$

where $E(x_i | y_i, z_i, w_i, \beta_{x \cdot uw}^{(d)})$ is the conditional mean of x_i given (y_i, z_i, w_i) , the values of (Y, Z, W) for unit i , $\sigma_{xx \cdot uw}^{(d)}$ is the residual variance of the distribution of X given U and W , and $e_i^{(d)}$ is a draw from the standard normal distribution. This method is proper in the sense discussed in Chapter 10, because it takes into account uncertainty in estimating the parameters.

The external calibration data are not generally available in the postimputation analysis. Reiter (2008) shows that in this situation, the standard MI combining rules in Chapter 10 yield a positively biased estimate of sampling variance, and resulting confidence interval coverage exceeds the nominal rate. Reiter (2008) describes an alternative two-stage imputation procedure to generate imputations that lead to consistent estimation of sampling variances. Specifically, we first draw d values of model parameters $\phi^{(d)}$; second, for each $\phi^{(d)}$, $d = 1, \dots, m$, we construct n imputed data sets by generating n sets of draws of X . The resulting $m \times n$ imputed datasets are then analyzed by the following combining rules:

For $d = 1, \dots, m$ and $i = 1, \dots, n$, let $\hat{\gamma}^{(d,i)}$ and $\text{Var}(\hat{\gamma}^{(d,i)})$ be the estimated parameters of interest and the associated estimated sampling variance computed with $D^{(d,i)}$ data set, respectively. The MI estimate of γ , γ_{MI} , and associated sampling variance T_{MI} are calculated as

$$\hat{\gamma}_{\text{MI}} = \sum_{d=1}^m \bar{\gamma}_n^{(d)}, \quad \text{where } \bar{\gamma}_n^{(d)} = \sum_{i=1}^n \hat{\gamma}^{(d,i)} / (mn),$$

$$T_{\text{MI}} = U - W + (1 + 1/m)B - W/n,$$

where

$$W = \sum_{d=1}^m \sum_{i=1}^n \left(\hat{\gamma}^{(d,i)} - \bar{\gamma}_n^{(d)} \right)^2 / (m(n-1)),$$

$$B = \sum_{d=1}^m \left(\bar{\gamma}_n^{(d)} - \hat{\gamma}_{\text{MI}} \right) / (m-1),$$

$$U = \sum_{d=1}^m \sum_{i=1}^n \text{Var}(\hat{\gamma}^{(d,i)}) / (mn).$$

The 95% interval for γ is $\hat{\gamma}_{\text{MI}} \pm t_{0.975, v} \sqrt{T_{\text{MI}}}$, with degrees of freedom

$$v = \left[\frac{((1+1/m)B)^2}{(m-1)T_{\text{MI}}} + \frac{((1+1/n)W)^2}{\% (m(n-1)) T_{\text{MI}}} \right]^{-1}.$$

When $T_{\text{MI}} < 0$, the sampling variance estimator is recalculated as $(1+1/m)B$, and inferences are based on a t -distribution with $(m-1)$ degrees of freedom.

Problems

- 11.1 Show that the available-case estimates of the means and variances of an incomplete multivariate sample, discussed in Section 3.4, are ML when the data are specified as multivariate normal with unrestricted means and variances, and zero correlations, with ignorable nonresponse. (This result implies that the available-cases method works reasonably well when the correlations are low.)
- 11.2 Write a computer program for the EM algorithm for bivariate normal data with an arbitrary pattern of missing values.
- 11.3 Write a computer program for generating draws from the posterior distribution of the parameters, for bivariate normal data with an arbitrary pattern of missing values, and a noninformative prior for the parameters.
- 11.4 Describe the EM algorithm for bivariate normal data with means (μ_1, μ_2) , correlation ρ , and common variance σ^2 , and an arbitrary pattern of missing values. If you did Problem 11.2, modify the program you wrote to handle this model. (*Hint:* For the M step, transform to $U_1 = Y_1 + Y_2$, $U_2 = Y_1 - Y_2$.)

- 11.5** Derive the expression for the expected information matrix in Section 11.2.2, for the special case of bivariate data.
- 11.6** For bivariate data, find the ML estimate of the correlation ρ for (a) a bivariate sample of size r , with known means (μ_1, μ_2) and known variances (σ_1^2, σ_2^2) , and (b) a bivariate sample of size r , and effectively infinite supplemental samples from the marginal distributions of both variables. Note the rather surprising fact that (a) and (b) yield different answers.
- 11.7** Prove the statement before Eq. (11.9) that complete-data ML estimates of Σ are obtained from C by simple averaging. (*Hint:* Consider the covariance matrix of the four variables $U_1 = Y_1 + Y_2 + Y_3 + Y_4$, $U_2 = Y_1 - Y_2 + Y_3 - Y_4$, $U_3 = Y_1 - Y_3$, and $U_4 = Y_2 - Y_4$.)
- 11.8** Review the discussion in Rubin and Thayer (1978, 1982) and Bentler and Tanaka (1983) on EM for factor analysis.
- 11.9** Derive the EM algorithm for the model of Example 11.4 extended with the specification that $\mu \sim N(0, \tau^2)$, where μ is treated as missing data. Then consider the case where $\tau^2 \rightarrow \infty$, yielding a flat prior on μ .
- 11.10** Examine Beale and Little's (1975) approximate method for estimating the covariance matrix of estimated slopes in Section 11.4.2, for a single predictor X , and data with (a) Y completely observed and X subject to missing values, and (b) X completely observed and Y subject to missing values. Does the method produce the correct asymptotic covariance matrix in either case?
- 11.11** Fill in the details leading to the expressions for the mean and variance of y_{j+1} given y_j, y_{j+2} , and θ in Example 11.8. Comment on the form of the expected values of y_{j+1} as $\beta \uparrow 1$ and $\beta \downarrow 0$.
- 11.12** For Example 11.8, extend the results of Problem 11.11 to compute the means, variances, and covariance of y_{j+1} and y_{j+2} given y_j, y_{j+3} and θ , for a sequence where y_j and y_{j+3} are observed, and y_{j+1} and y_{j+2} are missing.
- 11.13** Develop a Gibbs' sampler for simulating the posterior distributions of the parameters and predictions of the $\{z_i\}$ for Example 11.9. Compare the posterior distributions for the predictions for years 1949 and 1981 with the EM predictions in the last two columns of Table 11.6.

12

Models for Robust Estimation

12.1 Introduction

A feature of the model-based approach to statistics in general is that it makes explicit assumptions about the distribution of the data. Some view this as a negative, arguing that all models are wrong,¹ so inference based on models is always flawed. On the other hand, other approaches to inference, such as methods based on generalized estimating equations, often have implicit assumptions that are rarely subject to scrutiny. Making assumptions explicit lays them open to criticism, and criticism can lead to fixing doubtful assumptions and generating better inferences.

Robust estimation, and more generally robust inference, concerns methods that do not rely on strong assumptions about the structure of the model or the form of the error distribution. The early literature (see for example Andrews et al. 1972; Hampel et al. 1986) focused mainly on methods for estimating the center of a *symmetric* distribution that are resistant to outliers. This literature was not primarily model-based, but models can be formulated that also yield inferences that are resistant to outliers. These models generally replace the normal distribution for modeling continuous data by a longer-tailed distribution like the *t*-distribution, or skewed extensions of these distributions. In particular, Examples 8.4, 8.8–8.10, and 10.4 concerned robust inference from a single sample based on the *t*-distribution, with degrees of freedom fixed a priori or estimated from the data. In Section 12.2, we develop this idea more generally by considering alternative distributions to the normal (including the *t*-distribution) for robust inference, and robust inference for multivariate data sets with missing values. As in Chapter 11, we assume that the missingness mechanism is ignorable, and that categorical variables are present only in the form of fixed, fully-observed covariates. Robust inference for problems involving mixtures of continuous and categorical variables is considered in Chapter 14.

Section 12.3 considers a different form of robustness, namely to the form of the relationship of the mean of an outcome variable to regressor variables. In particular, with a continuous outcome Y and single predictor X , simple linear regression assumes a linear relationship between Y and X . One might seek to make this model more robust by assuming a more flexible shape for the relationship between the mean of Y and X . Polynomial regression is one way of doing this, but a more flexible approach is to fit a spline function that assumes distinct piecewise polynomials over regions of X delimited by values of X called knots, where these polynomials are constrained to provide some degree of smoothness at the knots, where the smoothness is specified by a penalty function, defined later. With missing data, a robust form of modeling is achieved by applying a spline model relating an incomplete variable Y to a constructed X , namely the propensity for Y to be missing, estimated as a function of the covariates. We call prediction of missing Y s based on this model penalized spline of propensity prediction (PSPP), and it is a useful alternative to the weighting methods discussed in Chapter 3. A key feature of this method is the use of the propensity to respond (that is, to be not missing) as a predictor, rather than as the basis for weighting.

12.2 Reducing the Influence of Outliers by Replacing the Normal Distribution by a Longer-Tailed Distribution

12.2.1 Estimation for a Univariate Sample

Dempster et al. (1977, 1980) consider maximum likelihood (ML) estimation for the following model, which generally leads to a nonnormal marginal distribution for the observed data. Let $X = (x_1, \dots, x_n)^T$ be an independent random sample from a population such that

$$(x_i | w_i, \theta) \sim_{\text{ind}} N(\mu, \sigma^2/w_i),$$

where $\{w_i\}$ is unobserved independent, identically distributed (iid) positive scalar random variable with known density $h(w_i)$. Inferences about $\theta = (\mu, \sigma)^T$ can be based on incomplete data methods, treating $W = (w_1, \dots, w_n)^T$ as missing data.

If X and W were both observed, then ML estimates of (μ, σ) would be found by weighted least squares:

$$\hat{\mu} = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i = s_1/s_0, \quad (12.1)$$

$$\hat{\sigma}^2 = \sum_{i=1}^n w_i (x_i - \hat{\mu})^2 / n = (s_2 - s_1^2/s_0) / n, \quad (12.2)$$

where $s_0 = \sum_{i=1}^n w_i$, $s_1 = \sum_{i=1}^n w_i x_i$, and $s_2 = \sum_{i=1}^n w_i x_i^2$ are the complete-data sufficient statistics for θ as defined in Section 8.4.2. When W is not observed, the ML estimates can be found by the expectation–maximization (EM) algorithm, treating the weights as missing data. The $(t + 1)$ th iteration of EM is as follows:

E Step

Estimate s_0 , s_1 , and s_2 by their conditional expectations, given X and current estimates $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)2})$ of θ . Because s_0 , s_1 , and s_2 are linear in the w_i 's, the E step reduces to finding estimated weights

$$w_i^{(t)} = E(w_i | x_i, \mu^{(t)}, \sigma^{(t)2}). \quad (12.3)$$

M Step

Compute new estimates $(\mu^{(t+1)}, \sigma^{(t+1)2})$ from (12.1) and (12.2), with (s_0, s_1, s_2) replaced by their estimates from the E step, that is, with w_i replaced by $w_i^{(t)}$ from (12.3). Convergence can be speeded by replacing the denominator n in (12.2) by $\sum_{i=1}^n w_i^{(t)}$, as in the parameter-expanded expectation–maximization (PX-EM) algorithm of Example 8.10.

This EM algorithm is iteratively reweighted least squares, with weights (12.3) that depend on the assumed distribution for w_i . Examples 8.4, 8.8–8.10, and 10.3 described the case where w_i is a scaled chi-squared distribution. Another choice, useful when the data contain extreme outliers, is given by the following example:

Example 12.1 *The Univariate Contaminated Normal Model.* Suppose that $h(w_i)$ is positive at two values of w_i , 1 and $\lambda < 1$, such that

$$h(w_i) = \begin{cases} 1 - \pi, & \text{if } w_i = 1, \\ \pi, & \text{if } w_i = \lambda, \text{ known,} \\ 0, & \text{otherwise,} \end{cases} \quad (12.4)$$

where $0 < \pi < 1$. Then the marginal distribution of x_i is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, \sigma^2/\lambda)$, which is a contaminated normal model, with probability of contamination π . For example, $\lambda = 0.1$ if the contamination is assumed to inflate the variance of x_i by a factor of 10.

A simple application of Bayes' theorem yields

$$E(w_i | x_i, \mu, \sigma^2) = \frac{1 - \pi + \pi \lambda^{3/2} \exp \{(1 - \lambda)d_i^2/2\}}{1 - \pi + \pi \lambda^{1/2} \exp \{(1 - \lambda)d_i^2/2\}} \quad (12.5)$$

and

$$\Pr(w_i = \lambda) = 1 - \Pr(w_i = 1) = \frac{\pi \lambda^{1/2} \exp\{(1-\lambda)d_i^2/2\}}{1 - \pi + \lambda^{1/2} \exp\{(1-\lambda)d_i^2/2\}}, \quad (12.6)$$

where

$$d_i^2 = (x_i - \mu)^2 / \sigma^2. \quad (12.7)$$

The weights $w_i^{(t)}$ for the t th iteration of EM are obtained by substituting current estimates $\mu^{(t)}$ and $\sigma^{(t)}$ in (12.5)–(12.7). Observe that values of x_i far from the mean have large values of d_i^2 and hence (for $\lambda < 1$) reduced weights in the M step. Thus, the algorithm leads to a robust estimate of μ , in the sense that outliers are downweighted.

Data augmentation for simulating the posterior distribution of the parameters under this model is similarly straightforward: the E step of EM is replaced by an I step that draws $w_i = 1$ or $w_i = \lambda$ with probabilities given by (12.6), computed at current draws of the parameters. The M step is replaced by a P step that draws new parameters from the complete-data posterior distribution for a normal sample, with observations weighted according to the previous I step. The draws are obtained as a special case of Example 6.16 with regressors confined to the constant term.

A straightforward and important practical extension of the models in Examples 8.4 and 12.1 is to model the mean as a linear combination of predictors X , yielding a weighted least squares algorithm for linear regression with contaminated normal or t errors (Rubin 1983a). Pettitt (1985) describes ML estimation for the contaminated normal and t models when the values of X are grouped and rounded.

12.2.2 Robust Estimation of the Mean and Covariance Matrix with Complete Data

Rubin (1983a) generalizes the model of Section 12.2 to multivariate data and applies it to derive ML estimates for contaminated multivariate normal and multivariate t samples. Let x_i be a $(1 \times K)$ vector of values of variables X_1, \dots, X_K . Suppose that x_i has the K -variate normal distribution

$$(x_i | \theta, w_i) \sim_{\text{ind}} N_K(\mu, \Psi/w_i), \quad (12.8)$$

where $\{w_i\}$ are unobserved iid positive scalar random variables with known density $h(w_i)$. ML estimates of μ and Ψ can be found by applying the EM algorithm, treating $W = (w_1, \dots, w_n)^T$ as missing data.

If W were observed, ML estimates of μ and Ψ would be the multivariate analogs of (12.1) and (12.2). The complete-data sufficient statistics for μ and

Ψ are $s_0 = \sum_{i=1}^n w_i$, $s_1 = \sum_{i=1}^n w_i x_i$, and $s_2 = \sum_{i=1}^n w_i x_i^T x_i$, and

$$\hat{\mu} = s_1/s_0 = \sum_{i=1}^n w_i x_i / \sum_{i=1}^n w_i \quad (12.9)$$

$$\hat{\Psi} = \frac{s_2 - s_1^T s_1 / s_0}{n} = \sum_{i=1}^n \frac{w_i (x_i - \hat{\mu})^T (w_i - \hat{\mu})}{n}. \quad (12.10)$$

Hence, when W is not observed, the $(t+1)$ th iteration of EM is as follows:

E Step

Estimate s_0 , s_1 , and s_2 by their conditional expectations, given X and current estimates $(\mu^{(t)}, \Psi^{(t)})$ of the parameters. Because s_0 , s_1 , and s_2 are linear in w_i , the E step reduces to finding estimated weights

$$w_i^{(t)} = E(w_i | x_i, \mu^{(t)}, \Psi^{(t)}).$$

M Step

Compute new estimates $(\mu^{(t+1)}, \Psi^{(t+1)})$ from (12.9) and (12.10), with s_0 , s_1 , and s_2 replaced by their estimates from the E step. The algorithm is speeded by replacing the denominator n in (12.9) by the sum of the current weights, $\sum_{i=1}^n w_i^{(t)}$, which corresponds to a PX-EM algorithm, as discussed in Section 8.5.3.

When the $\{w_i\}$ are distributed as (12.4), the marginal distribution of x_i is a mixture of $N(\mu, \Psi)$ and $N(\mu, \Psi/\lambda)$, that is, we obtain a contaminated K -variate normal model. The weights are then given by the following generalizations of (12.5)–(12.7):

$$E(w_i | x_i, \mu, \Psi) = \frac{1 - \pi + \pi \lambda^{K/2+1} \exp\{(1-\lambda)d_i^2/2\}}{1 - \pi + \pi \lambda^{K/2} \exp\{(1-\lambda)d_i^2/2\}}, \quad (12.11)$$

where d_i^2 is now the squared Mahalanobis distance for unit i :

$$d_i^2 = (x_i - \mu)^T \Psi^{-1} (x_i - \mu). \quad (12.12)$$

The model downweights units with large values of d_i^2 , as in the univariate situation. If, on the other hand, the distribution of $w_i \sim_{\text{ind}} \chi_v^2/v$, the weights are given by the following generalization of (8.24):

$$E(w_i | x_i, \mu, \Psi) = (\nu + K) / (\nu + d_i^2), \quad (12.13)$$

where d_i^2 is again given by (12.12).

Rubin (1983a) also considers extensions of these models to robust multivariate regression.

12.2.3 Robust Estimation of the Mean and Covariance Matrix from Data with Missing Values

Little (1988b) extends these robust algorithms to situations where some values of multivariate X are missing. Let $x_{(0), i}$ denote the set of variables observed for unit i , let $x_{(1), i}$ denote the missing variables, and write $X_{(0)} = \{x_{(0), i} : i = 1, \dots, n\}$ and $X_{(1)} = \{x_{(1), i} : i = 1, \dots, n\}$. We assume that first, x_i given w_i has the distribution given by (12.8) and second, the missing data are MAR. ML estimates of μ and Ψ are found by applying the EM algorithm, treating the values of both $X_{(1)}$ and W as missing data.

The M step here is identical to the M step when X is completely observed, described in the previous section. The E step estimates the complete-data sufficient statistics s_0 , s_1 , and s_2 by their conditional expectations, given $X_{(0)}$ and current estimate $\theta^{(t)} = (\mu^{(t)}, \Psi^{(t)})$ of the parameter θ . We find

$$E(s_0 | X_{(0)}) = E\left(\sum_{i=1}^n w_i \mid \theta^{(t)}, x_{(0),i}\right) = \sum_{i=1}^n w_i^{(t)},$$

where $w_i^{(t)} = E(w_i | \theta^{(t)}, x_{(0),i})$; the j th component of $E(s_1 | \theta^{(t)}, X_{(0)})$ is

$$\begin{aligned} E\left(\sum_{i=1}^n w_i x_{ij} \mid \theta^{(t)}, X_{(0)}\right) &= \sum_{i=1}^n E\{w_i E(x_{ij} | \theta^{(t)}, x_{(0),i}, w_i) | \theta^{(t)}, x_{(0),i}\} \\ &= \sum_{i=1}^n w_i^{(t)} \hat{x}_{ij}^{(t)}, \end{aligned}$$

where $\hat{x}_{ij}^{(t)} = E(x_{ij} | \theta^{(t)}, x_{(0),i})$, because the conditional mean of x_{ij} given $(\theta^{(t)}, x_{(0),i}, w_i)$ does not depend on w_i . Finally, the (j,k) th element of $E(s_2 | \theta^{(t)}, X_{(0)})$ is

$$\begin{aligned} E\left(\sum_{i=1}^n w_i x_{ij} x_{ik} \mid \theta^{(t)}, X_{(0)}\right) &= \sum_{i=1}^n E\{w_i E(x_{ij} x_{ik} | \theta^{(t)}, x_{(0),i}, w_i) | \theta^{(t)}, x_{(0),i}\} \\ &= \sum_{i=1}^n \left(w_i^{(t)} \hat{x}_{ij}^{(t)} \hat{x}_{ik}^{(t)} + \psi_{jk(0),i}^{(t)} \right), \end{aligned}$$

where the adjustment $\psi_{jk(0),i}^{(t)}$ is zero if x_{ij} or x_{ik} is observed, and w_i times the residual covariance of x_{ij} and x_{ik} given $x_{(0),i}$, if x_{ij} and x_{ik} are both missing.

The quantities $\hat{x}_{ij}^{(t)}$ and $\psi_{jk(0),i}^{(t)}$ are found by sweeping the current estimate $\Psi^{(t)}$ of Ψ to make $x_{(0),i}$ predictor variables, computations identical to those of the normal EM algorithm (Section 11.2.1). The only modification needed to the latter algorithm is to weight, by $w_i^{(t)}$, the sums and sums of squares and cross products used in the next M step.

The weights $w_i^{(t)}$ for the contaminated normal and t models are simple modifications of the weights when data are complete: they are given by Eqs. (12.11) and (12.13), respectively, with (i) K replaced by K_i , the number of observed variables for unit i , and (ii) the squared Mahalanobis distance (12.12) computed using only the observed variables for unit i .

Both the multivariate t and contaminated normal models downweight units with large squared distances, d_i^2 . The distribution of the weights, however, is quite different for the two models, as the following example illustrates.

Example 12.2 *Distribution of Weights from Multivariate t and Contaminated Multivariate Normal Models.* Figure 12.1 shows the distribution of weights for (a) the multivariate t with $v = 6.0$, (b) the multivariate t with $v = 4.0$, and (c) the contaminated normal model with $\pi = 0.1$, $\lambda = 0.077$, for multivariate t_4 data with $K = 4$ variables, $n = 80$ units, and 72 of the 320 values randomly deleted. Observe that the weights are more dispersed for $v = 4$ than for $v = 6$, and the downweighting for the contaminated normal model tends to be concentrated in a few outlying units.

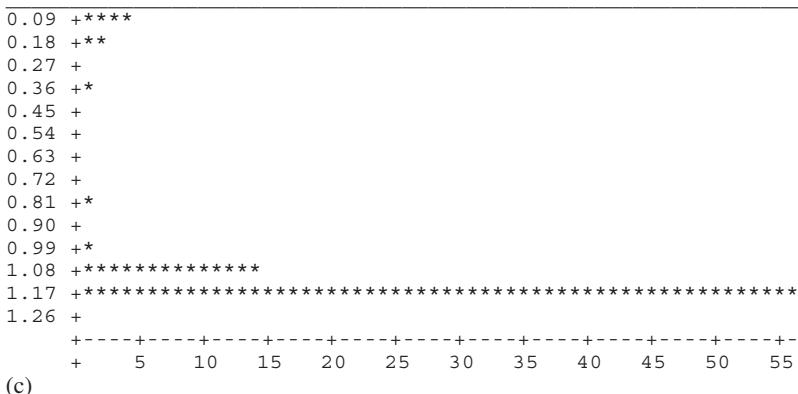
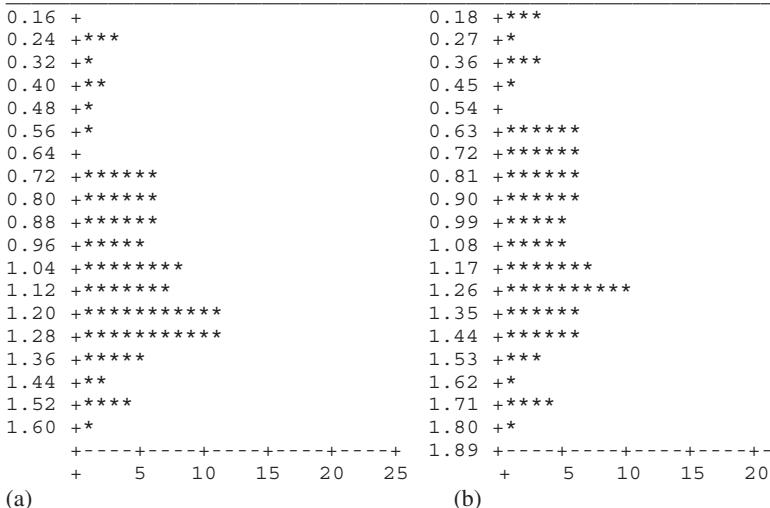
Little (1988b) shows, by a simulation study, that ML for the models in Example 12.2 can produce estimates of means, slopes, and correlations that are protected against outliers when the data are nonnormal, with minor sacrifices of efficiency when the data are, in fact, normal.

12.2.4 Adaptive Robust Multivariate Estimation

The methods discussed so far assume that the parameters v of the t model or (π, λ) of the contaminated normal model are known. Estimating these parameters in addition to the mean and scale parameters yields a form of adaptive robust estimation. ML estimation of v for the univariate t model was described in Examples 8.8 and 8.9. These methods are readily extended to provide adaptive robust ML estimates for the multivariate problems in Sections 12.2.2 and 12.2.3. The t th iteration of an ECME algorithm (as discussed in Section 8.5.2) for the multivariate incomplete data of Section 12.2.3 follows:

E step: Take the expectation of the complete-data sufficient statistics, as in Section 12.2.3, with v replaced by current estimate $v^{(t)}$.

CM Steps: Compute new estimates $(\mu^{(t+1)}, \psi^{(t+1)})$ as in Sections 12.2.2 and 12.2.3. Compute $v^{(t+1)}$ to maximize the observed-data loglikelihood $\ell(\mu^{(t+1)}, \Psi^{(t+1)}, v | X_{(0)})$ with respect to v . This is a one-dimensional maximization and can be achieved by a grid search or Newton stepping.



*Weights are scaled to average to one.

Figure 12.1 Example 12.2, distributions of weights from robust ML methods, applied to multivariate t_4 data. (a) Multivariate t_6 model; (b) multivariate t_4 model; and (c) contaminated normal model.

12.2.5 Bayes Inference for the t Model

The methods of ML estimation in Sections 12.2.1–12.2.3 can be quite easily modified to yield draws from the posterior distribution of the parameters. Consider in particular the multivariate t model for x_i , assuming the relatively diffuse prior distribution on v :

$$p(\mu, \Psi, \ln(1/v)) \propto |\Psi|^{-(K+1)/2}, -10 < \ln(1/v) < 10, \quad (12.14)$$

which is used in Liu and Rubin (1998). Let $(\mu^{(t)}, \Psi^{(t)}, v^{(t)})$ and $(x_{(1),i}^{(t)}, w_i^{(t)}, i = 1, \dots, n)$ be draws of the parameters and missing values at iteration t . Iteration $t+1$ consists of the following computations:

- (a) For $i = 1, \dots, n$, draw new weights $w_i^{(t+1)} \sim u_i / (v^{(t)} + d_{(0),i}^{2(t)})$, where $u_i \sim \chi^2_{v^{(t)} + K_i}$.
- (b) For $i = 1, \dots, n$, draw missing data $x_{(1),i}^{(t+1)} = \hat{x}_{(1),i}^{(t)} + z_i$, where $\hat{x}_{(1),i}^{(t)}$ is the predicted mean from the (normal) linear regression of $x_{(1),i}$ on $x_{(0),i}$ given current parameter estimates, and z_i is normal with mean 0 and covariance matrix $(w_i^{(t)})^{-1} \Psi_{(1)\cdot(0),i}^{(t)}$, where $\Psi_{(1)\cdot(0),i}^{(t)}$ is obtained by sweeping $\Psi^{(t)}$ on the observed variables $x_{(0),i}$ for unit i .
- (c) Draw $(\mu^{(t+1)}, \Psi^{(t+1)})$ from the posterior distribution of these parameters given filled-in data $(X_{(0)}, X_{(1)}^{(t+1)})$ and weights $W^{(t+1)}$. This is a standard complete-data problem, with $\Psi^{(t+1)}$ being drawn from a scaled inverse-Wishart distribution, and $\mu^{(t+1)}$ given $\Psi^{(t+1)}$ being drawn from a K -variate normal distribution.
- (d) Draw $v^{(t+1)}$ from its posterior distribution, given the current parameters, filled-in data and weights. This posterior distribution does not have a simple form, but because v is scalar, obtaining a draw is not difficult, and can be accomplished in a number of ways. For the computations shown here, the griddy Gibbs' sampler (see, for example Tanner 1996, section 6.4) was employed, with 400 equally spaced cut-points on the interval $(-10, 10)$.

Convergence of the algorithm can be speeded by defining the missing values to yield a monotone pattern, and then developing the P step for a monotone missing-data pattern, using the ideas of Chapter 7. This approach is described in Liu (1995). A straightforward extension is to robust multivariate regression with fully observed covariates, where the fixed covariates are swept in the weighted scale matrix augmented by a column of means (Liu 1996). These methods are applied in the next example.

Example 12.3 Robust MANOVA with Missing Data Illustrated Using the St. Louis Data (Example 11.6 Continued). A Bayesian analysis for the MANOVA model with multivariate t errors and prior (12.14) was fitted to the data in Table 11.1, including an indicator variable for the low and medium/high-risk groups as in Example 11.6. The regression coefficient of the group indicator measures the difference in mean outcome between the medium/high and low-risk groups. Figure 12.2 displays draws of these estimated mean differences for the four outcomes, and 95% posterior probability intervals are shown below the histograms. The intervals tend to be slightly narrower than intervals (Figure 11.2) under the normal model of Example 11.6, although conclusions are

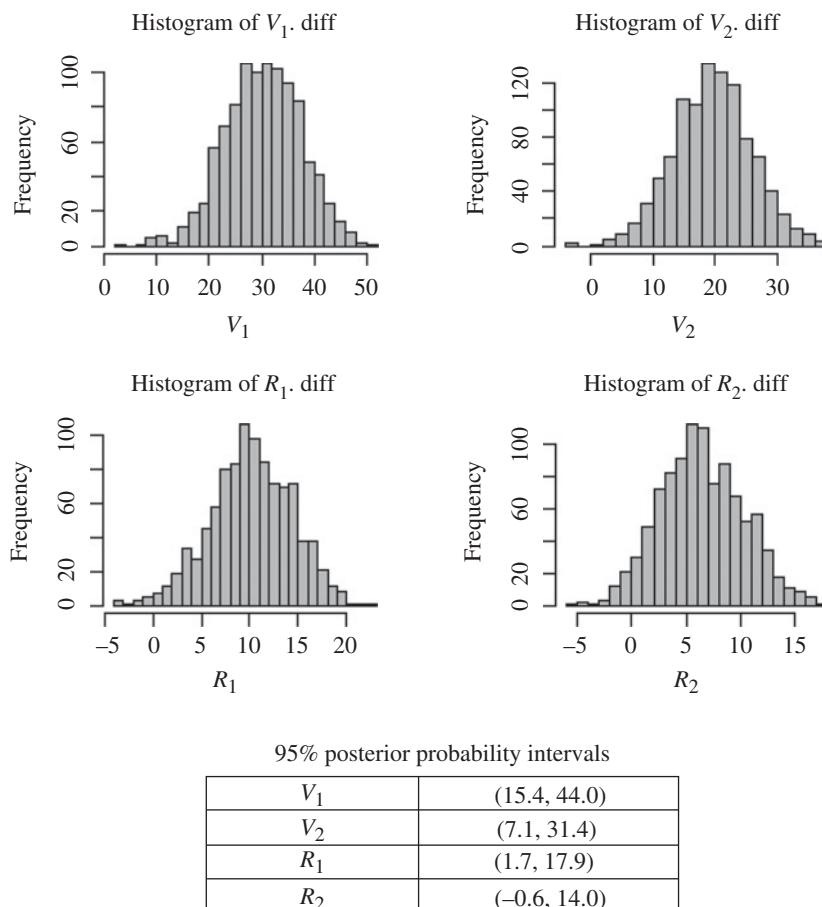


Figure 12.2 Example 12.3, posterior distributions of mean differences $\mu_{\text{low}} - \mu_{\text{med/high}}$, St. Louis risk research data, based on 9000 draws, multivariate t regression model.

similar. The posterior distribution of v is quite dispersed and sensitive to the choice of prior distribution for v , but inferences for the differences in means are less sensitive to this modeling choice.

12.2.6 Further Extensions of the t Model

The next example generalizes the multivariate t model by allowing constraints on the mean and covariance matrix, as in the normal model of Section 11.5.

Example 12.4 *Robust ML Estimation of Repeated Lung-Function Measures with Missing Values.* Lange et al. (1989) analyze data reported by LaVange

(1983) from a longitudinal study of lung function conducted on 72 children aged 3–12 years at a child development center. The variables consist of race (black or white), sex, and up to eight yearly measurements of $\log(v_{\max_{75}})$ for each child, where $v_{\max_{75}}$ is the maximum expiratory flow rate after 75% of the forced vital capacity has been exhaled. Of the ten annual $v_{\max_{75}}$ measures for each child from age 3 to 12, the number actually recorded ranges from 1 to 8, with an average of 4.3 per child; thus the amount of missing data is substantial. Some measures of $v_{\max_{75}}$ at early and late ages (for example ages 3 and 12) are never observed together, given that the study was only eight years long. Hence, the full covariance matrix of $v_{\max_{75}}$ measurements for ages 3–12 is not uniquely estimable without placing restrictions on the parameters.

The results in Table 12.1 show whether there are differences in the growth curves of $\log(v_{\max_{75}})$ over time between males and females. Let y_{ij} denote the value of $\log(v_{\max_{75}})$ for individual i at age $j + 2$ for $1 \leq j \leq 10$. Table 12.1(a) shows the maximized loglikelihoods for normal repeated-measures models of the form (11.20), namely

$$y_i \sim_{\text{ind}} N_{10}(\mu_i(\beta), \Sigma(\psi)). \quad (12.15)$$

The covariance matrix Σ is modeled as $\sigma_{jk} = \psi_1(\psi_2 + (1 - \psi_2)\psi_3^{|k-j|})$, where ψ_1 determines the total dispersion, ψ_2 is a heritability parameter, and ψ_3 is an environmental decay constant. The j th component of the mean $\mu_i(\beta)$ has the form:

$$\mu_{ij} = \begin{cases} \beta_0 + \beta_1 \text{age}_j + \beta_2 \text{age}_j^2, & \text{if } \text{sex}_i = \text{male}, \\ \beta_3 + \beta_4 \text{age}_j + \beta_5 \text{age}_j^2, & \text{if } \text{sex}_i = \text{female}, \end{cases} \quad (12.16)$$

thereby modeling distinct quadratic curves relating lung function to age among males and females. The quadratic terms model nonlinearity in the absence of any theory-based functional form for the curves. Overall, the full model (labeled 1N in the table) has nine parameters, six for the mean function and three for the covariance matrix. Table 12.1(a) shows the maximized loglikelihood and likelihood-ratio chi-squared statistics for models 2N to 6N that place the indicated restrictions on the parameters. The model 5N appears to be the best-fitting parsimonious normal model.

Table 12.1(b) shows fits for the same set of models with the normal distribution in (12.15) replaced by the t , with degrees of freedom v estimated from the data by ML. Note that these models fit much better than their normal counterparts, with maximized loglikelihoods 15–23 larger than the maximized loglikelihoods for the corresponding normal models. Based on the t models, (4T) seems a reasonable summary of the data. That is, the lung-function curves appear linear, with no differences between males and females. Interestingly, the model 5T that sets the heritability parameter equal to 0 does not fit well, unlike

Table 12.1 Example 12.4, normal models of lung function data

(a) Normal models					
Model	Mean constraints	Covariance constraints	Number of parameters	Maximized loglikelihood	Likelihood ratio (LR) statistic compared to model 1N (df)
1N	None	None	9	164.4	—
2N	$\beta_2 = \beta_5 = 0$	None	7	164.1	0.7 (2)
3N	$\beta_2 = \beta_5 = 0,$ $\beta_1 = \beta_4$	None	6	161.9	5.2 (3)
4N	$\beta_2 = \beta_5 = 0,$ $\beta_0 = \beta_3, \beta_1 = \beta_4$	None	5	161.4	6.2 (4)
5N	None	$\psi_3 = 0$	8	163.5	2.0 (1)
6N	None	$\psi_2 = 0$	8	156.4	16.5 ^a (1)

(b) T models					
Model	Mean constraints	Covariance constraints	Number of parameters	Maximized loglikelihood	LR statistic compared to model 1T (df)
1T	None	None	10	187.1	—
2T	$\beta_2 = \beta_5 = 0$	None	8	186.2	2.0 (2)
3T	$\beta_2 = \beta_5 = 0,$ $\beta_1 = \beta_4$	None	7	184.8	4.7 (3)
4T	$\beta_2 = \beta_5 = 0,$ $\beta_0 = \beta_3, \beta_1 = \beta_4$	None	6	184.2	5.9 (4)
5T	None	$\psi_3 = 0$	9	183.1	8.1 ^a (1)
6T	None	$\psi_2 = 0$	9	181.5	9.2 ^a (1)

Summary of fits for 12 Models.

^aSignificantly worse fit than (a) Model 1N or (b) 1T at the 1% level (likelihood ratio chi-squared test).

the corresponding normal model 5N. It appears that for the normal model, outliers are obscuring the (expected) decline in the covariances as the time between measurements increases. Parameter estimates from the best-fitting *t* model 4T and the corresponding normal model 4N are shown in Table 12.2 with asymptotic standard errors based on a numerical approximation of the observed information matrix. Note that the best-fitting *t* has between 4 and 5 df and

Table 12.2 Example 12.4, normal models of lung function data

Model	β_0	β_1	ψ_1	ψ_2	ψ_3	ν
4N	-0.365 (0.075)	0.0637 (0.0102)	0.167 (0.017)	0.362 (0.076)	0.175 (0.092)	∞ (-)
4T	-0.286 (0.069)	0.0608 (0.0090)	0.109 (0.017)	0.406 (0.092)	0.304 (0.102)	4.4 (1.2)

Parameter estimates and standard errors for models 4N and 4T.

increases the size and statistical significance of the ψ_3 , parameter, as expected from the model comparisons in Table 12.1. The slopes of the regression lines are similar for the normal and t fits, but the intercept for the t fit is noticeably smaller.

A limiting feature of the models described here is that the scaling quantity q_i applied to models with longer-than-normal tails is the same for all the variables in the data set. It may be desirable to allow different scaling factors for different variables, reflecting, for example, different degrees of contamination. In particular, for robust regression with missing predictors, it may be more appropriate to confine the scaling quantity to the outcome variable.

Unfortunately, if the models are extended to allow different scaling factors for different variables, the simplicity of the E step of the EM algorithm is lost for general patterns of missing data. Some exceptions worth mentioning are based on the fact that the models can be readily extended to handle a set of fully observed covariates Z , such that y_i has a multivariate normal linear regression on z_i with mean $\sum \beta_j z_{ij}$ and covariance matrix Ψ/q_i , conditional on the unknown scaling quantity q_i defined as before. Thus, suppose the data can be arranged in a *monotone* missingness pattern, with the variables arranged in blocks X_1, \dots, X_k such that for $j = 1, \dots, K-1$, X_j is observed for all units where X_{j+1} is observed. Then the joint distribution of X_1, \dots, X_k can be expressed as the product of distributions

$$f(X_1, \dots, X_K | \phi) = f(X_1 | \phi) f(X_2 | X_1, \phi) \cdots f(X_K | X_1, \dots, X_{K-1}, \phi),$$

as discussed in Chapter 7. The conditional distribution $f(X_j | X_1, \dots, X_{j-1})$ in this factorization can then be modeled as multivariate normal with mean $\sum_{u=1}^{j-1} \beta_u X_u$ and covariance matrix $\Psi_{j,1\dots j-1}/q_{jj}$, where now the scaling factors q_{jj} vary for different values of j . The parameters of each component of the likelihood are estimated by the multivariate regression generalization of the model just considered, and then ML estimates of other parameters of the joint distribution of X_1, \dots, X_k are found by transformation, as discussed in Chapter 7.

12.3 Penalized Spline of Propensity Prediction

We consider the following missing-data problem. Let (Y, X_1, \dots, X_p) be a vector of variables with Y observed for units $i = 1, \dots, r$ and missing for $i = r+1, \dots, n$, with fully observed covariates X_1, \dots, X_p . We consider estimation and inference for the mean of Y , under the assumption that the missingness of Y depends only on X_1, \dots, X_p , so the missingness mechanism is MAR.

Propensity weighting, as discussed in Chapter 3, is one possible approach to this problem. However, this method can yield estimates with large sampling variances when respondents with very small propensity scores are assigned excessively large weights. Also, weighted complete-case analysis does not fully exploit information on the covariates in the incomplete units. To address this problem, we can multiply-impute the missing values with draws from their predictive distribution under a model for the distribution of Y given X_1, \dots, X_p . Inferences for the parameters can be derived using the combining rules described in Chapter 10. This MI approach is efficient if the imputation model is well specified, but is potentially vulnerable to misspecification of the imputation model. This motivates PSPP (Little and An 2004; Zhang and Little 2009) which bases imputations on a spline model that avoids restrictive model assumptions about the relationship between the mean of Y and X_1, \dots, X_p .

Define the logit of the propensity for Y to be observed as

$$P^*(\phi) = \text{logit}(\Pr(M = 0 \mid X_1, \dots, X_p), \phi). \quad (12.17)$$

PSPP imputations are predictions from the following model:

$$(Y \mid P^*(\phi), X_1, \dots, X_p \mid \theta, \beta, \phi) \sim N(s(P^*(\phi) \mid \theta) + g(P^*(\phi), X_2, \dots, X_p \mid \beta), \sigma^2), \quad (12.18)$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and constant variance σ^2 . The first component of the mean function, $s(P^*(\phi) \mid \theta)$, is a spline function of the propensity score P^* , indexed by parameters θ . The second component $g(P^*(\phi), X_2, \dots, X_p \mid \beta)$ is a parametric function, which includes any covariates other than P^* that predict Y . One of the predictors, here X_1 , is omitted from the g function to avoid possible multicollinearity.

One choice for $s(P^*(\phi) \mid \theta)$ is a penalized spline (Eilers and Marx 1996; Ruppert et al. 2003) of the form

$$s(P^*(\phi) \mid \theta) = \theta_0 + \theta_1 P^* + \sum_{k=1}^K \gamma_k (P^* - \kappa_k)_+, \quad (12.19)$$

where $1, P^*, (P^* - \kappa_1)_+, \dots, (P^* - \kappa_K)_+$ is the truncated linear basis; $\kappa_1 < \dots < \kappa_K$ are a priori selected fixed knots, and K is the total number of knots; and $(\gamma_1, \dots, \gamma_K)$ are random effects assumed to be normally distributed with mean 0 and variance τ^2 . ML estimates for this model can be found using a number of

existing software packages, such as PROC MIXED in SAS (SAS 1992; Ngo and Wand 2004) and lme in S-plus (Pinheiro and Bates 2000); for Bayesian methods using Winbugs, see Crainiceanu et al. (2005). The first step of fitting a PSPP model estimates the propensity score, for example by a logistic regression model or probit model of M on X_1, \dots, X_p ; the second step fits the regression of Y on estimated P^* as a spline model with the other covariates included in the model parametrically in the g function. When Y is a continuous variable, we choose a normal distribution with constant variance. For other types of data, extensions of the PSPP can be formulated by using the generalized linear models with different link functions.

Three approaches to fitting the PSPP model are (i) ML, where parameters are estimated by ML and standard errors are computed using the information matrix or the bootstrap; (ii) Bayes, where parameters are drawn from their posterior distribution and inference about μ is based on draws from its posterior distribution; and (iii) multiple imputation (MI) (PSPP-MI), where draws of the missing values are multiply imputed, and inferences based on Rubin's MI combining rules.

Parameter uncertainty is propagated here by drawing parameters for each completed data set from their posterior distribution, or estimating the parameters on a bootstrap sample of the original data, assuming asymptotic normality holds (Heitjan and Little 1991). Zhang and Little (2011) used PSPP-MI with the bootstrap to represent parameter uncertainty; the Bayesian version of PSPP with dispersed prior distributions for the parameters may be preferable in small samples.

Alternative approaches to PSPP use the inverse of the propensity as a weight. In particular, augmented inverse probability-weighted (AIPW) estimates (Robins and Rotnitzky 1995; Robins et al. 1995) have the form:

$$\hat{\mu} = n^{-1} \left(\sum_{i=1}^n \hat{y}_i \right) + n^{-1} \left(\sum_{i=1}^r \hat{w}_i (y_i - \hat{y}_i) \right),$$

where $\hat{w}_i = 1/\hat{P}(m_i = 0 | X_1, \dots, X_p)$ is the estimated weight for the i th subject, and \hat{y}_i is the prediction from a parametric model for the i th subject. This method has a so-called "double robustness" (DR) property, in the sense that it yields consistent estimates of the mean if either the prediction model for Y or the propensity model is correctly specified. Bang and Robins (2005) propose a regression-based method that also has the DR property.

The PSPP method has a form of DR as well, deriving from the balancing property of the propensity score. The latter states that, under MAR and correct specification of the propensity model (12.17), the distribution of covariates conditional on the propensity is the same for respondents and nonrespondents to Y (Rosenbaum and Rubin 1983). Consequently, the average of the observed and imputed values of Y is consistent if either (a) the mean of Y given (P^*, X_1, \dots, X_p) in model (3) is correctly specified, or (b1) the propensity P^* is correctly

specified, and (b2) $E(Y|P^*) = s(P^*)$, implying that the regression function g does not have to be correctly specified (Little and An 2004; Zhang and Little 2009). Condition (b2) is arguably a weak assumption, because the relationship of the mean of Y with the propensity is modeled flexibly by a spline function.

We note that the DR definition differs somewhat from the original concept of robustness, which refers to resistance of procedures to violations of model assumptions; simulations in Kang and Schafer (2007) suggest that the DR property is less useful when both prediction and propensity models are mildly misspecified. Simulations in Zhang and Little (2011) and Yang and Little (2015), including situations similar to those considered by Kang and Schafer, suggest that PSPP compares favorably with AIPW and other DR approaches.

Problems

- 12.1** Derive the weighting function (12.5) for the model of Example 12.1.
- 12.2** Describe the data augmentation algorithm for Example 12.1.
- 12.3** Outline a program to compute ML estimates for the contaminated normal model of Example 12.4. Simulate data from the contaminated normal model and explore sensitivity of inferences to different choices of true and assumed values of π and λ .
- 12.4** Explore ML estimation for the contaminated normal model of Example 12.1 with (a) π known and λ unknown and estimated by ML, and (b) π unknown and estimated by ML and λ known. (Does the case with both π and λ unknown involve too much missing information to be practical?)
- 12.5** Extend the t model with known degrees of freedom and the contaminated normal to the case of simple linear regression of X on a fixed observed covariate Z . Derive the EM algorithm for this model. Do units with Z observed, but X missing, contribute information? (*Hint:* Review Section 11.4.1.)
- 12.6** Derive the weighting functions (12.11) and (12.13) for the models of Section 12.2.2.
- 12.7** Derive the E step equations in Section 12.2.3.

Note

¹ In George Box's famous phrase, "all models are wrong, but some are useful."

13

Models for Partially Classified Contingency Tables, Ignoring the Missingness Mechanism

13.1 Introduction

This chapter concerns the analysis of incomplete data when variables are categorical. Although interval-scaled variables can be handled by forming categories based on segments of the scale, the ordering between the categories of variables treated in this way, or of other ordinal variables, is not exploited in the methods considered here. However, methods for categorical data that take into account orderings between categories (e.g., Goodman 1979; McCullagh 1980) could be extended to handle incomplete data, by applying the likelihood theory of Part II.

A rectangular $(n \times V)$ data matrix consisting of n units on V categorical variables Y_1, \dots, Y_V can be rearranged as a V -dimensional contingency table, with C cells defined by joint levels of the variables. The entries in the table are counts $\{n_{j_k \dots u}\}$, where $n_{j_k \dots u}$ is the number of sampled units in the cell with $Y_1 = j, Y_2 = k, \dots, Y_V = u$. If the data matrix has missing items, some of the units in the preceding contingency table are partially classified. The completely classified units create a V -dimensional table of counts $\{r_{j_k \dots u}\}$, and the incompletely classified units create supplemental lower-dimensional sub-tables, each defined by the subset of variables (Y_1, \dots, Y_v) that are observed. For example, the first eight rows of Table 1.2 provide data from the complete units in a five-way contingency table with variables Sex, Age Group, and Obesity at three time points. The remaining 18 rows provide data on the six partially classified tables with one or two of the obesity variables missing. We discuss ML and Bayes estimation for data of this form.

In the next section, factorizations of the likelihood analogous to those discussed in Chapter 7 for normal data are applied to special patterns of incomplete categorical data. Estimation for general patterns of missingness using the EM algorithm and posterior simulation is discussed in Section 13.3.

Section 13.4 considers ML and Bayes estimation for partially classified data when the classification probabilities are constrained by a loglinear model. Non-ignorable nonresponse models for categorical data are deferred until Section 15.4.2.

A more general type of incomplete data occurs when level j of a particular variable, Y_1 , say, is not known, but it is known that the unit falls into one of a subset S of values of Y_1 . If Y_1 is completely missing, then S consists of all the possible values of Y_1 . If Y_1 is missing but the value of a less detailed recode Y_1^* of Y_1 is recorded, then S will be a proper subset of the possible values of Y_1 . An example of such data subject to coarse and refined classifications is given in Example 13.4.

The missing-data situations considered here should be carefully distinguished from the situations of “structural zeros,” where certain cells contain zero counts because the model assigns them zero probability of containing any entry. For example, if Y_1 = year of birth and Y_2 = year of first marriage, marriages before birth are impossible, and cells with $Y_2 \leq Y_1$ are “structural zeros” in the joint distribution of Y_1 and Y_2 . For discussion of the structural zero problem, see, for example Bishop et al. (1975, chapter 5).

13.2 Factored Likelihoods for Monotone Multinomial Data

13.2.1 Introduction

In this section, we assume that the complete-data counts $\{n_{jk\dots u}\}$ have a multinomial distribution with total count n and probabilities $\theta = \{\pi_{jk\dots u}\}$. We also assume that the missingness mechanism is missing at random (MAR), in the sense discussed in Chapter 6, and that the missing-data pattern is monotone. Thus, the likelihood for the probabilities θ is obtained by integrating the complete-data likelihood

$$L(\theta \mid \{n_{jk\dots u}\}) = \prod_{j,k,\dots,u} \pi_{jk\dots u}^{n_{jk\dots u}} , \quad \sum_{j,k,\dots,u} \pi_{jk,\dots,u} = 1, \quad (13.1)$$

over the missing data. ML estimates of θ are obtained by maximizing the resulting likelihood, subject to the constraint that the cell probabilities sum to 1.

An alternative to the multinomial model assumes that the cell counts $\{n_{jk\dots u}\}$ are independent Poisson random variables with means $\{\mu_{jk\dots u}\}$ and cell probabilities $\pi_{jk\dots u}^* = \mu_{jk\dots u} / \sum_{j,k,\dots,u} \mu_{jk\dots u}$. If the missingness mechanism is MAR, likelihood inferences for $\{\pi_{jk\dots u}^*\}$ are the same as those for $\{\pi_{jk\dots u}\}$ under the multinomial model. This fact follows from arguments analogous to those for the complete-data case (Bishop et al. 1975). We restrict attention to the multinomial model because it seems more common than the Poisson model in practical situations.

For complete data, the likelihood (13.1) yields the ML estimate

$$\hat{\pi}_{jk\cdots u} = n_{jk\cdots u}/n.$$

With asymptotic sampling variance

$$\text{Var}(\hat{\pi}_{jk\cdots u} | \theta)|_{\theta=\hat{\theta}} = \hat{\pi}_{jk\cdots u}(1 - \hat{\pi}_{jk\cdots u})/n.$$

For Bayes' inference, we multiply the complete-data likelihood by the conjugate Dirichlet prior for the cell probabilities:

$$p(\{\pi_{jk\cdots u}\}) \propto \prod_{j,k,\dots,u} \pi_{jk\cdots u}^{\alpha_{jk\cdots u}-1}, \quad \pi_{jk\cdots u} > 0, \quad \sum_{j,k,\dots,u} \pi_{jk\cdots u} = 1. \quad (13.2)$$

Combining this prior distribution with the likelihood yields the Dirichlet posterior distribution

$$p(\{\pi_{jk\cdots u}\} | \{n_{jk\cdots u}\}) \propto \prod_{j,k,\dots,u} \pi_{jk\cdots u}^{\alpha_{jk\cdots u} + n_{jk\cdots u} - 1}, \quad \sum_{j,k,\dots,u} \pi_{jk\cdots u} = 1 \quad (13.3)$$

(see Example 6.18).

Our objective is to obtain analogous answers from incomplete data. In this section, we discuss parameterizations for special patterns of incomplete data that yield explicit ML estimates and direct simulation of the posterior distribution.

13.2.2 ML and Bayes for Monotone Patterns

We first consider ML estimation for a two-way contingency table with one supplemental margin.

Example 13.1 *Two-Way Contingency Table with One Supplemental One-Way Margin.* Consider two categorical variables Y_1 , with levels $j = 1, \dots, J$ and Y_2 , with levels $k = 1, \dots, K$. The data consist of r units $(y_{i1}, y_{i2}, i = 1, \dots, r)$ with y_{i1} and y_{i2} observed and $m = n - r$ units $(y_{i1}, i = r + 1, \dots, n)$ with y_{i1} observed and y_{i2} missing. The data pattern is identical to Example 7.1, but the variables are now categorical.

The r completely classified units can be displayed in a $(J \times K)$ contingency table, with r_{jk} units in the cell with $y_{i1} = j$, $y_{i2} = k$. The $n - r$ remaining units form a supplemental $(J \times 1)$ margin, with m_j units in the cell with $y_{i1} = j$ (see Figure 13.1).

		Y_2		Total			
Levels		1	\dots	K	Total	Levels	
Y_1 :	1			$\{r_{j+}\}$	r	1	$\{m_j\}$
	J			$\{r_{+k}\}$		J	
Total					Total		

Complete units Incomplete units

Figure 13.1 The data in Example 13.1.

We shall use the standard “plus” notation for summation over subscripts j and k . For this problem,

$$\theta = (\pi_{11}, \pi_{12}, \dots, \pi_{JK}) \quad \text{and} \quad \sum_{j=1}^J \sum_{k=1}^K \pi_{jk} \equiv \pi_{++} = 1.$$

By analogy with Example 7.1, we adopt the alternative parameter set ϕ corresponding to the marginal distribution of Y_1 and the conditional distribution of Y_2 given Y_1 . The likelihood of the data can be written

$$L(\phi | \{r_{jk}, m_j\}) = \left(\prod_{j=1}^J \pi_{j+}^{r_{j+} + m_j} \right) \times \left(\prod_{j=1}^J \prod_{k=1}^K \pi_{k,j}^{r_{jk}} \right), \quad (13.4)$$

where the first factor is the likelihood for the multinomial distribution of the marginal counts $r_{j+} + m_j$, with total count n and probabilities π_{j+} , and the second factor is the likelihood for the product of J conditional multinomial distributions of $\{r_{jk}\}$ given r_{j+} , with total count r_{j+} and probabilities

$$\pi_{k,j} = \Pr(Y_2 = k | Y_1 = j) = \pi_{jk} / \pi_{j+}, \quad k = 1, \dots, K.$$

Note that (13.4) is a factorization of the likelihood as discussed in Section 7.1, with distinct parameters

$$\phi_1 = \{\pi_{j+}, j = 1, \dots, J\} \quad \text{and} \quad \phi_2 = \{\pi_{k,j}, j = 1, \dots, J; k = 1, \dots, K\}.$$

Maximizing each factor in (13.4) separately, we obtain ML estimates

$$\hat{\pi}_{j+} = \frac{r_{j+} + m_j}{n}, \quad \hat{\pi}_{k,j} = \frac{r_{jk}}{r_{j+}},$$

so

$$\hat{\pi}_{jk} = \hat{\pi}_{j+} \hat{\pi}_{k,j} = \frac{[r_{jk} + (r_{jk}/r_{j+})m_j]}{n}. \quad (13.5)$$

These ML estimates effectively distribute a proportion r_{jk}/r_{j+} of the unclassified units m_j into the (j, k) th cell.

For a Bayesian analysis, we specify for simplicity independent Dirichlet prior distributions for $\{\pi_{j+}\}$ and $\{\pi_{k\cdot}\}$, corresponding to the factored likelihood in (13.2):

$$p(\phi) \propto \left(\prod_{j=1}^J \pi_{j+}^{n_{j0}-1} \right) \times \left(\prod_{j=1}^J \prod_{k=1}^K \pi_{k\cdot j}^{r_{jk0}-1} \right).$$

The posterior distribution is then a product of independent posterior distributions for $\{\pi_{j+}\}$ and $\{\pi_{k\cdot j}\}$, namely:

$$p(\phi | \text{data}) \propto \left(\prod_{j=1}^J \pi_{j+}^{n_{j0} + r_{j+} + m_j - 1} \right) \times \left(\prod_{j=1}^J \prod_{k=1}^K \pi_{k\cdot j}^{r_{jk} + r_{jk0} - 1} \right). \quad (13.6)$$

A draw $\pi_{jk}^{(d)}$ of π_{jk} from its posterior distribution is easily obtained by first drawing $\pi_{j+}^{(d)}$ and $\pi_{k\cdot j}^{(d)}$ from their independent posterior distributions in (13.6), and then setting $\pi_{jk}^{(d)} = \pi_{j+}^{(d)} \pi_{k\cdot j}^{(d)}$, the analog of (13.5). Drawing from a Dirichlet distribution is easily accomplished using chi-squared or gamma random variables, as described in Example 6.18.

Example 13.2 Numerical Illustration of ML and Bayes for Monotone Bivariate Count Data. A numerical illustration of the results of Example 13.1 is provided by the data in Table 13.1, where Y_1 is a dichotomous variable and Y_2 is a trichotomous one. The marginal probabilities of Y_1 are estimated from completely and partially classified units:

$$\hat{\pi}_{1+} = 190/410, \quad \hat{\pi}_{2+} = 220/410.$$

Table 13.1 Numerical example of the data pattern of Figure 13.1

			Complete units			Incomplete units		
Y_1	Y_2							
	1 2 3			Total				
	1	20	30	40	90	Y_1	1	100
	2	50	60	20	130		2	90
Total			70 90 60	220		Total	190	

The conditional probabilities of classification for Y_2 given Y_1 are estimated from the completely classified units:

$$\hat{\pi}_{1.1} = 20/90, \quad \hat{\pi}_{2.1} = 30/90, \quad \hat{\pi}_{3.1} = 40/90,$$

$$\hat{\pi}_{1.2} = 50/130, \quad \hat{\pi}_{2.2} = 60/130, \quad \hat{\pi}_{3.2} = 20/130.$$

Hence, the estimated probabilities (13.5) are the following:

$$\hat{\pi}_{11} = (20/90)(190/410) = 0.1030, \quad \hat{\pi}_{21} = (50/130)(220/410) = 0.2064,$$

$$\hat{\pi}_{12} = (30/90)(190/410) = 0.1545, \quad \hat{\pi}_{22} = (60/130)(220/410) = 0.2377,$$

$$\hat{\pi}_{13} = (40/90)(190/410) = 0.2060, \quad \hat{\pi}_{23} = (20/130)(220/410) = 0.0826.$$

In contrast, estimates based only on the completely classified units are as follows:

$$\tilde{\pi}_{11} = 20/220 = 0.0909, \quad \tilde{\pi}_{21} = 50/220 = 0.2273,$$

$$\tilde{\pi}_{12} = 30/220 = 0.1364, \quad \tilde{\pi}_{22} = 60/220 = 0.2727,$$

$$\tilde{\pi}_{13} = 40/220 = 0.1818, \quad \tilde{\pi}_{23} = 20/220 = 0.0909.$$

Under MAR, the estimates $\{\hat{\pi}_{jk}\}$ are less efficient than the ML estimates $\{\tilde{\pi}_{jk}\}$. However, as in the normal case discussed in Example 7.1, the principal value of ML is its ability to reduce or eliminate bias when the data are MAR but not missing completely at random (MCAR). The estimates $\{\hat{\pi}_{jk}\}$ are ML if the data are MAR, and in particular if the probability that Y_2 is missing depends on Y_1 , but not Y_2 . The $\{\tilde{\pi}_{jk}\}$ are consistent for $\{\pi_{jk}\}$ in general only if the data are MCAR, that is, missingness does not depend on the values of Y_1 and Y_2 . Because the marginal distribution of Y_1 appears to be different for the completely and incompletely classified samples (a chi-squared test yields $\chi^2_1 = 5.23$, with associated p value <0.01), the MCAR assumption appears to be implausible; in this example, there can be no evidence against MAR.

For a Bayesian analysis of this example, suppose we assume the following independent Jeffreys' prior distributions for $\{\pi_{j+}\}$ and $\{\pi_{k+j}\}$:

$$p(\phi) \propto \left(\pi_{1+}^{-1/2} \pi_{2+}^{-1/2} \right) \left(\pi_{1.1}^{-1/2} \pi_{2.1}^{-1/2} \pi_{3.1}^{-1/2} \right) \left(\pi_{1.2}^{-1/2} \pi_{2.2}^{-1/2} \pi_{3.2}^{-1/2} \right).$$

Then the posterior distribution of these parameters is also a product of Dirichlet distributions:

$$p(\phi | \text{data}) \propto (\pi_{1+}^{189.5} \pi_{2+}^{219.5}) (\pi_{1.1}^{19.5} \pi_{2.1}^{29.5} \pi_{3.1}^{39.5}) (\pi_{1.2}^{49.5} \pi_{2.2}^{59.5} \pi_{3.2}^{19.5}).$$

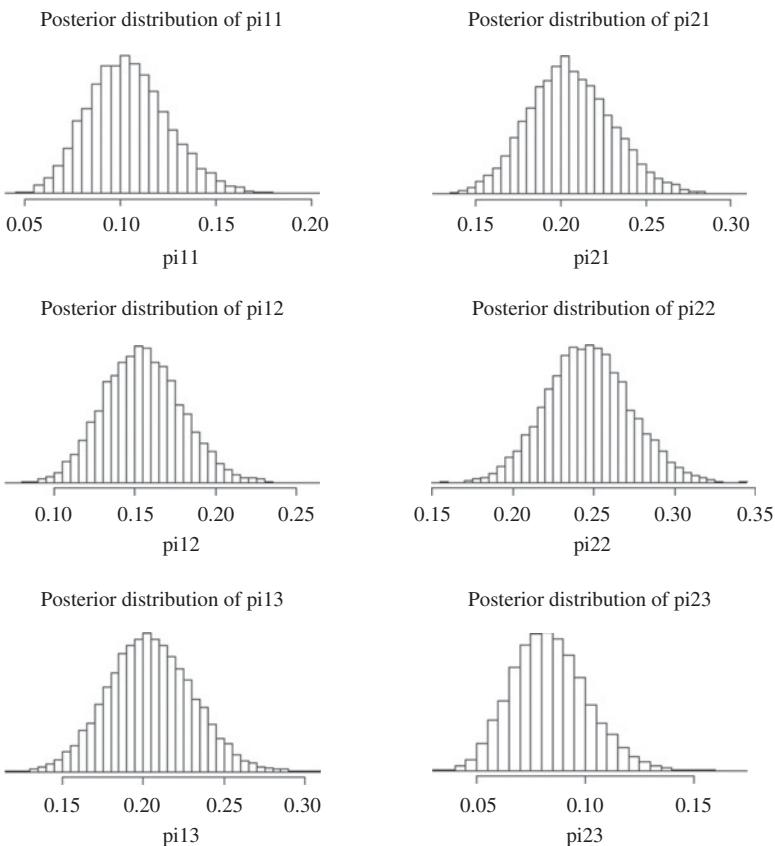


Figure 13.2 Example 13.2, plots of posterior distributions of the cell probabilities.

Drawing from these distributions and setting $\pi_{jk}^{(d)} = \pi_{j+}^{(d)} \pi_{k+j}^{(d)}$ for $j = 1, 2$ and $k = 1, 2, 3$ yields the posterior distributions displayed in Figure 13.2. The posterior means are similar to the ML estimates:

$$E(\pi_{11} | \text{data}) = 0.1044, \quad E(\pi_{21} | \text{data}) = 0.2058,$$

$$E(\pi_{12} | \text{data}) = 0.1549, \quad E(\pi_{22} | \text{data}) = 0.2467,$$

$$E(\pi_{13} | \text{data}) = 0.2045, \quad E(\pi_{23} | \text{data}) = 0.0838.$$

A useful feature of the Bayesian analysis is that it yields estimates of precision from the simulated posterior distributions. These are discussed in Example 13.5 below.

Extensions of this example to other monotone patterns can be developed by analogous factorizations of the likelihood.

Table 13.2 Partially classified contingency table for Example 13.3

Mental	Physical	Survival	Sex = Male				Sex = Female			
			Age = <75		Age = >75		Age = <75		Age = >75	
			E ^a	C ^a	E	C	E	C	E	C
<i>(a) Fully categorized</i>										
Poor	Poor	Deceased	0	2	5	3	0	0	2	1
		Survived	1	0	0	0	0	0	0	1
	Good	Deceased	0	0	2	2	1	1	1	0
		Survived	0	2	2	0	0	0	0	0
Good	Poor	Deceased	0	0	3	1	0	0	1	2
		Survived	3	1	1	2	0	1	1	0
	Good	Deceased	1	1	4	6	2	0	0	2
		Survived	5	10	6	8	3	5	2	4
<i>(b) Missing physical status</i>										
Poor	Missing	Deceased	0	0	0	0	0	0	0	0
		Survived	0	0	1	0	0	0	0	0
Good		Deceased	0	0	0	0	0	0	0	0
		Survived	0	0	0	0	0	0	0	0
<i>(c) Missing mental status</i>										
Missing	Poor	Deceased	2	0	5	3	1	1	2	0
		Survived	1	1	0	3	0	0	0	1
	Good	Deceased	1	0	0	0	0	0	1	2
		Survived	1	3	2	1	1	1	0	0
<i>(d) Missing both physical and mental status</i>										
Missing	Missing	Deceased	0	1	2	2	1	0	3	1
		Survived	2	8	1	2	1	1	2	2

^aE denotes experimental group. C denotes control group. D = survival status, G = group, S = sex, A = age, P = physical status, M = mental status.

Example 13.3 Application to a Six-Way Table. Fuchs (1982) presents data from the Protective Services Project for Older Persons, a longitudinal study of 164 people designed to assess the effect of enriched social casework services on the well being of clients (Table 13.2). Investigators collected data on six dichotomized variables, D = survival status (survived, deceased), G = group membership (experimental, control), S = sex (male, female), A = age (less than

75, over 75), P = physical status (poor, good), M = mental status (poor, good). The data on all the variables were available on 101 participants (Table 13.2(a)). Physical status was missing for one participant (Table 13.2(b)). Mental status was missing for 33 participants (Table 13.2(c)). Finally, physical and mental status were both missing on 29 participants (Table 13.2(d)).

When the information on mental status for the single unit in Table 13.2(b) is ignored, the data have a monotone pattern, and ML estimates of the cell probabilities can be derived using the factorization

$$\begin{aligned} \Pr(D, G, A, S, P, M | \theta) &= \Pr(D, G, A, S | \theta) \Pr(P | D, G, A, S, \theta) \\ &\quad \times \Pr(M | D, G, A, S, P, \theta). \end{aligned}$$

The observed counts for estimating the three distributions on the right side are displayed in Table 13.3(a). The resultant expected cell counts, which are the estimated cell probabilities multiplied by the total sample size, 164, are displayed in Table 13.3(b); for example, the count for the cell with D = survived, G = experimental, A = over 75, S = male, P = good, M = good is

$$164 \left(\frac{13}{164} \right) \left(\frac{10}{11} \right) \left(\frac{6}{8} \right) = 8.8636.$$

Changing D = survived to D = deceased yields the expected count

$$164 \left(\frac{21}{164} \right) \left(\frac{6}{19} \right) \left(\frac{4}{6} \right) = 4.4211.$$

Hence, the estimated conditional probability of survival given G = experimental, A = over 75, S = male, P = good, M = good is $(8.8636 / (8.8636 + 4.4211)) = 0.6672$. This estimate compares with $10 / (10 + 6) = 0.6$ for the complete units in Table 13.2.

Example 13.4 *Tables with Refined and Coarse Classifications.* The data in Table 13.4, presented and analyzed by Hocking and Oxspiring (1974), illustrate another situation where ML and Bayes estimates can be found by factoring the likelihood. Table 13.4(a) gives data on the use of drugs in the treatment of leprosy. The 196 patients are classified according to the degree of infiltration and overall clinical condition, after a fixed time over which treatments were administered. The supplemental data in Table 13.4(b) on 400 patients are classified coarsely with respect to improvement in health. Such data arise naturally in health surveys where detailed results can be obtained for a small group of individuals, and coarsely classified data can be collected inexpensively for a larger group of individuals.

The likelihood factors according to the joint distribution of the combined cell counts from the two tables, classified coarsely as in Table 13.4(b), based on all 596 patients, and the conditional distribution of degree of improvement

Table 13.3 Example 13.3, ML estimation for monotone data^a in Table 13.2(a), (c), and (d), using factored likelihood method

Mental	Physical	Survival	Experimental	Control	Male		Female					
					<75		>75					
					Experimental	Control	Experimental	Control				
<i>(a) Partitioned table for monotone patterns</i>												
(i) Available information on D, G, A, S												
Deceased	4	4	21	17	5	2	10	8				
Survived	13	25	13	16	5	8	5	8				
(ii) Available information on D, G, A, S, P												
Poor	Deceased	2	2	13	7	1	1	5				
Poor	Survived	5	2	1	5	0	1	1				
Good	Deceased	2	1	6	8	3	1	2				
Good	Survived	6	15	10	9	4	6	4				
<i>(b) Expected cell frequencies</i>												
(iii) Available information on all variables (D, G, A, S, P, M) given in Table 13.2(a)												
Poor	Poor	Deceased	1.00	2.67	8.98	5.95	0.63	0.50				
Poor	Survived	1.48	0.00	0.00	0.00	0.00	0.00	0.00				
Good	Deceased	0.00	0.00	2.21	2.27	1.25	1.00	2.86				
Good	Survived	0.00	3.68	2.95	0.00	0.00	0.00	0.00				
Good	Poor	Deceased	1.00	0.00	5.39	1.98	0.63	0.50				
Good	Survived	4.43	2.94	1.18	5.71	0.00	1.14	1.67				
Good	Deceased	2.00	1.33	4.42	6.80	2.50	0.00	0.00				
Good	Survived	7.09	18.38	8.86	10.29	5.00	6.86	3.33				
<i>M = mental status.</i>												

^aThe information on the mental status of the individual in Table 13.2(b) is ignored. D = survival status, G = group, S = sex, A = age, P = physical status, M = mental status.

Source: Fuchs (1982), with minor corrections. Reproduced with permission of Taylor and Francis.

Table 13.4 Example 13.4, patients classified by degree of infiltration and change of condition**(a) Finely classified data**

Degree of infiltration	Clinical change					Total
	Marked	Moderate	Slight	Stationary	Worse	
Little	11	27	42	53	11	144
Much	7	15	16	13	1	52
Total	18	42	58	66	12	196

(b) Coarsely classified data

Degree of infiltration	Clinical change			Total
	Improvement	Stationary	Worse	
Little	144	120	16	280
Much	92	24	4	120
Total	236	144	20	400

(c) ML estimates of cell probabilities from (a) and (b)

Degree of infiltration	Clinical change				
	Improvement				
Marked	Moderate	Slight	Stationary	Worse	
Little	(224/596)(11/80)	(224/596)(27/80)	(224/596)(42/80)	173/596	27/596
Much	(130/596)(7/38)	(130/596)(15/38)	(130/596)(16/38)	37/596	5/596

Source: Hocking and Oxspring (1974). Reproduced with permission of John Wiley and Sons.

(marked, moderate, or slight) given improvement and degree of infiltration, based on the smaller group of 196 patients. Resulting ML estimates of the cell probabilities are displayed in Table 13.4(c), in a form that illustrates the calculations. The joint probabilities of infiltration and coarsely classified clinical change are obtained by merging the data in (a) and (b), yielding the fractions in the last two columns and the first factors of the first three columns. The latter are multiplied by the conditional probabilities of degree of improvement,

calculated from the first three columns of (a). In particular, the top left corner entry is $\hat{\pi}_{11} = (224/596)(11/80) = 0.0517$, compared with $\tilde{\pi}_{11} = 11/196 = 0.0561$ from the finely classified data alone.

13.2.3 Precision of Estimation

The asymptotic covariance matrix associated with the ML estimates (13.5) can be obtained by calculating the information matrix for the parameters in the factored form of the likelihood, inverting this matrix, and then transforming to the original parameterization using the method outlined in Section 7.1. Alternatively, we can calculate these variances and covariances directly. For example, to calculate the large-sample variance of $\hat{\pi}_{jk} = \hat{\pi}_{j+}\hat{\pi}_{k+j}$ in Example 13.1, we write

$$\text{Var } \hat{\pi}_{jk} = E(\text{Var } \hat{\pi}_{jk} | \{n_{1+}\}) + \text{Var}(E(\hat{\pi}_{jk} | \{n_{1+}\})),$$

where $\{n_{1+}\}$ is the set of marginal counts of Y_1 . Hence,

$$\begin{aligned} \text{Var } \hat{\pi}_{jk} &= E\{\hat{\pi}_{j+}^2 \pi_{k+j}(1 - \pi_{k+j})/r_{j+}\} + \text{Var}\{\hat{\pi}_{j+}\pi_{k+j}\} \\ &= \pi_{j+}^2 \pi_{k+j}(1 - \pi_{k+j})/r_{j+} + \pi_{k+j}^2 \pi_{j+}(1 - \pi_{j+})/n, \end{aligned}$$

asymptotically to order $1/r_{j+}^2$. Some algebra yields

$$\text{Var } \hat{\pi}_{jk} \approx \frac{\pi_{jk}(1 - \pi_{jk})}{r} \left\{ 1 - \frac{\pi_{k+j} - \pi_{jk}}{1 - \pi_{jk}} \frac{n - r}{n} + c_j \frac{1 - \pi_{k+j}}{1 - \pi_{jk}} \right\},$$

where $c_j = r\pi_{j+}/r_{j+} - 1$. Substituting estimates of the parameters yields

$$\text{Var}(\pi_{jk} - \hat{\pi}_{jk}) \approx \frac{\hat{\pi}_{jk}(1 - \hat{\pi}_{jk})}{r} \left\{ 1 - \frac{\hat{\pi}_{k+j} - \hat{\pi}_{jk}}{1 - \hat{\pi}_{jk}} \frac{n - r}{n} + c_j \frac{1 - \hat{\pi}_{k+j}}{1 - \hat{\pi}_{jk}} \right\}. \quad (13.7)$$

The left side of (13.7) is written in a modified form to indicate that a Bayesian analysis of the asymptotic posterior variance of π_{jk} yields similar results. For the covariances, we find

$$\begin{aligned} \text{Cov}(\pi_{jk} - \hat{\pi}_{jk}, \pi_{jl} - \hat{\pi}_{jl}) &\approx \frac{-\hat{\pi}_{jk}\hat{\pi}_{jl}}{r} \left\{ 1 + \frac{(1 - \hat{\pi}_{j+})}{\hat{\pi}_{j+}} \frac{n - r}{n} + \frac{c_j}{\hat{\pi}_{j+}} \right\}, \quad k \neq l, \\ \text{Cov}(\pi_{ik} - \hat{\pi}_{ik}, \pi_{jl} - \hat{\pi}_{jl}) &\approx \frac{-\hat{\pi}_{ik}\hat{\pi}_{jl}}{n}, \quad i \neq j. \end{aligned}$$

For interval estimates, a generally more satisfactory approach, especially when the sample size is small, is to calculate asymptotic variances on a transformation of π_{jk} that better satisfies asymptotic normality, for example $\text{logit}(\pi_{jk})$.

An even better approach is to simulate the full posterior distribution of the $\{\pi_{jk}\}$, and form interval estimates from the central $100(1 - \alpha)\%$ of the drawn values.

Example 13.5 Estimates of Precision for Bivariate Monotone Multinomial Data (Example 13.2 Continued). We now compare the estimated precisions of the complete-case, ML and Bayes estimates of π_{11} from data in Table 13.1. If the data are MCAR, the complete-case estimate $\tilde{\pi}_{11} = 0.0909$, which ignores the supplementary margin, has large-sample variance $\pi_{11}(1 - \pi_{11})/r$, which after substituting ML estimates yields

$$\text{Var}(\tilde{\pi}_{11}) = \frac{(0.1030)(1 - 0.1030)}{220} = 0.000\,420. \quad (13.8)$$

Similarly, from (13.7), the ML estimate $\hat{\pi}_{11} = 0.1030$ has estimated large-sample variance

$$\text{Var}(\hat{\pi}_{11}) \approx 0.000\,42(0.9384 + 0.1151) = 0.000\,442. \quad (13.9)$$

Thus, the inclusion of the supplemental margin does not increase the estimated precision. However, as noted in Example 13.2, the data do not appear to be MCAR, so $\tilde{\pi}_{11}$ is probably not a consistent estimate of π_{11} , and so (13.8) is not a valid estimate of the precision of $\tilde{\pi}_{11}$ as an estimate of π_{11} . Assuming the missing data are MAR, $\hat{\pi}_{11}$ is consistent for π_{11} , so a rough estimate of the bias of $\tilde{\pi}_{11}$ is $\tilde{\pi}_{11} - \hat{\pi}_{11} = 0.0121$. Hence, a rough estimate of the mean squared error of $\tilde{\pi}_{11}$ is

$$\text{MSE}(\tilde{\pi}_{11}) \approx 0.0121^2 + \text{Var}(\tilde{\pi}_{11}) = 0.000\,566.$$

Comparing this with (13.9), the ML procedure appears considerably more precise when the bias of the complete-case estimate is taken into account.

The posterior distribution of π_{11} provides a better estimate of precision than these asymptotic results. The posterior variance of the distribution in Figure 13.1 is $\text{Var}(\pi_{11} | \text{data}) = 0.000\,444$, slightly larger than the asymptotic large-sample variance of the ML estimate. A central 95% posterior probability interval for π_{11} from the plot in Figure 13.1 reflects the skewness in this posterior distribution.

13.3 ML and Bayes Estimation for Multinomial Samples with General Patterns of Missingness

As with normal data, incomplete multinomial data that do not form a monotone data pattern require iterative methods for ML or Bayes estimation. The EM algorithm is particularly simple, because the loglikelihood is linear in the

missing values. For the monotone data in Examples 13.1 and 13.2, ML estimation effectively distributes the partially classified data into the full table, using conditional probabilities estimated from the fully classified data. The E step of the EM algorithm for general patterns has the same form, except that the conditional probabilities are calculated from current estimates of the cell probabilities rather than from the fully classified data. The M step of the EM algorithm calculates new cell probabilities from the data completed by the E step. This algorithm first appeared in the statistical literature in Hartley (1958). We provide a quite general formulation of the algorithm, and then apply it to a special case.

Suppose that the complete data are a multinomial sample of size n , with C cells, n_c units classified in cell c , and parameters $\theta = (\pi_1, \dots, \pi_C)$, where π_c is the classification probability for cell c . The observed data comprise r completely classified units, with r_c belonging in cell c for $c = 1, \dots, C$, and $r = \sum_{c=1}^C r_c$ and $n - r$ partially classified units, which belong to subsets of the C cells. For a multiway table with supplemental margins, the subsets consist of the cells that are aggregated to form each cell in the supplemental margins. We partition the partially classified units into K groups, within which all units have the same missingness pattern, that is, the same set of possible cells. Suppose that m_k partially classified units fall in the k th group, and let S_k denote the set of cells to which these units might belong. Furthermore, define the indicator functions $\delta(c \in S_k)$, $c = 1, \dots, C$, $k = 1, \dots, K$, where $\delta(c \in S_k) = 1$ if cell c belongs to S_k and $\delta(c \in S_k) = 0$ otherwise.

To define the E step of the EM algorithm, as earlier, let $\{\pi_c^{(t)}, c = 1, \dots, C\}$ denote the current (t th iterate) estimate of the parameters. The complete data belong to the regular exponential family, with a complete-data loglikelihood that is linear in the sufficient statistics

$$\{n_c, c = 1, \dots, C\}.$$

Hence the E step calculates

$$n_c^{(t)} = E \left\{ n_c \mid \text{data}, \pi_1^{(t)}, \dots, \pi_C^{(t)} \right\} = r_c + \sum_{k=1}^K m_k \psi_{c \cdot S_k}^{(t)},$$

where

$$\psi_{c \cdot S_k}^{(t)} = \pi_c^{(t)} \delta(c \in S_k) / \left(\sum_{j=1}^C \pi_j^{(t)} \delta(j \in S_k) \right)$$

is the current estimate of the conditional probability of falling in cell j given that a unit falls in the set of categories S_k . The E step effectively distributes the partially classified units into the table according to these probabilities.

Table 13.5 Example 13.6, a 2×2 table with supplemental margins for both variables

(1) Classified by Y_1 and Y_2			(2) Classified by Y_1			(3) Classified by Y_2		
	Y_2						Y_2	
	1	2	Total				1	2
Y_1	1	100	50	150	Y_1	1	30 ^a	
	2	75	75	150		2	60 ^b	
Total	175	125	300	Total		90		88

The superscripts a, b, c, d refer to the partially classified cells and are used in Table 13.6.

The M step calculates new parameter estimates as

$$\pi_c^{(t+1)} = n_c^{(t)} / n.$$

A Bayesian analysis is analogous, except that it uses the data augmentation (DA) algorithm to draw values of the missing data and the parameters. A simple numerical example illustrates this procedure:

Example 13.6 *A 2×2 Table with Supplemental Data on Both Margins.* ML estimation for two-way tables with supplementary data on both margins was first considered by Chen and Fienberg (1974). Table 13.5 gives data for a (2×2) table with supplemental margins for both the classifying variables, analyzed in Little (1982). Table 13.6 shows the first three iterations of the EM algorithm, where initially the cell probabilities are estimated from the completely classified table. These probabilities are then used to allocate the partially classified units. For example, the 28 partially classified units with $Y_2 = 1$ have $Y_1 = 1$ with probability $100/(100 + 75)$ and $Y_1 = 2$ with probability $75/(100 + 75)$. Thus of the 28 units, in effect $(28)(100)/175 = 16$ are allocated to $Y_1 = 1$ and $(28)(75)/175 = 12$ are allocated to $Y_1 = 2$. In the next step, new probabilities are found from the completed data and the procedure iterates to convergence. Final probabilities of classification after convergence are

$$\hat{\pi}_{11} = 0.28, \hat{\pi}_{12} = 0.17, \hat{\pi}_{21} = 0.24, \hat{\pi}_{22} = 0.31.$$

For a Bayesian analysis, assume again a Jeffreys' prior distribution for the parameters:

$$p(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \propto \pi_{11}^{-1/2} \pi_{12}^{-1/2} \pi_{21}^{-1/2} \pi_{22}^{-1/2}.$$

The I (imputation) step of DA draws missing values in the supplemental margins, based on current draws $(\pi_{11}^{(t)}, \pi_{12}^{(t)}, \pi_{21}^{(t)}, \pi_{22}^{(t)})$ of $(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. That is, the I step draws the 30 missing values of Y_2 with $Y_1 = 1$ in Part (2)

Table 13.6 Example 13.6, the EM algorithm for data in Table 13.5, ignoring the missingness mechanism

Estimated probabilities				Fractional allocation of units			
Step 1							
		Y_2				Y_2	
		1 2			1 2		
$\begin{array}{cc} Y_1 & \\ \begin{array}{cc} 1 & \\ 2 & \end{array} & \begin{array}{cc} 100/300 & 50/300 \\ 75/300 & 75/300 \end{array} \end{array}$				$\begin{array}{cc} Y_1 & \\ \begin{array}{cc} 1 & \\ 2 & \end{array} & \begin{array}{cc} 100 + 20^a + 16^c & 50 + 10^a + 24^d \\ 75 + 30^b + 12^c & 75 + 30^b + 36^d \end{array} \end{array}$			
				30^a 60^d			
Step 2				28^c 60^d			
$\begin{array}{cc} 136/478 & 84/478 \\ 117/478 & 141/478 \end{array}$				$\begin{array}{cc} 100 + 18.6 + 15.1 & 50 + 11.4 + 22.4 \\ 75 + 27.2 + 12.9 & 75 + 32.8 + 37.6 \end{array}$			
Step 3				$100 + 18.4 + 15.1$ $75 + 26.5 + 12.9$			
$\begin{array}{cc} 0.28 & 0.18 \\ 0.24 & 0.30 \end{array}$				$\begin{array}{cc} 50 + 11.6 + 21.9 \\ 75 + 33.5 + 38.1 \end{array}$			
Step 4							
$\begin{array}{cc} 0.28 & 0.17 \\ 0.24 & 0.31 \end{array}$							

The superscript in the top right panel indicate the partially classified cells in Table 13.5. For example, of the 28 units with $Y_2 = 1$ (superscript c), 16 are allocated to $Y_1 = 1$ and 12 are allocated to $Y_1 = 2$.

of Table 13.5 with $\Pr(Y_2 = 1 | Y_1 = 1, \pi^{(t)}) = \pi_{11}^{(t)} / \pi_{1+}^{(t)}$ and $\Pr(Y_2 = 2 | Y_1 = 1, \pi^{(t)}) = \pi_{12}^{(t)} / \pi_{1+}^{(t)}$; and analogously for the other partially classified counts in Parts (2) and (3) of the table.

The P (posterior) step of DA then draws new parameters $(\pi_{11}^{(t+1)}, \pi_{12}^{(t+1)}, \pi_{21}^{(t+1)}, \pi_{22}^{(t+1)})$ from the complete-data posterior distribution based on the filled-in data from the previous I step. Because this complete-data posterior distribution is Dirichlet, the method described in Example 6.17 can be applied to this step.

Example 13.7 Application of EM to Positron Emission Tomography. Vardi et al. (1985) applies the EM algorithm to two-way counted data from positron emission tomography (PET). The description here is from Rubin's (1985b) discussion. In PET, a "picture" of an organ (say, the brain) is created by collecting counts of emissions in D detectors placed systematically around the

organ. The organ is segmented into B boxes or pixels, each characterized by a distinct intensity parameter $\lambda(b)$, $b = 1, \dots, B$ governing the rate of emission. Physical considerations provide a $D \times B$ matrix of known conditional probabilities, $\Pr(\text{detector} = d | \text{pixel} = b)$, for the probability that an emission from pixel b will be recorded in detector d . The objective is to use these known conditional probabilities in conjunction with the observed counts in the D detectors to estimate the intensities (or marginal probabilities of emissions) in the B pixels.

Let $\pi = \{\pi(d, b)\}$ be the $D \times B$ matrix of joint probabilities that an emission emanates from pixel b and is detected in detector d ; π is determined by the $\Pr(d|b)$ and $\lambda(b)$. The complete data are n iid units, $\delta_i = \{\delta_i(d, b)\}$, where $\delta_i(d, b) = 1$ if the i th emission emanated from pixel b and is recorded in detector d and zero, otherwise. The observed (that is, incomplete) data comprise the n row margins of the δ_i , which are $D \times 1$ vectors indicating the detector for the n emissions. The EM algorithm proceeds as follows:

1. Start with some initial guess for λ , say $\lambda^{(0)}$, which implies an initial value for π , $\pi^{(0)}$.
2. At the E step, for $d = 1, \dots, D$, allocate the observed count in detector d across the B pixels according to the conditional probabilities implied by $\pi^{(0)}$.
3. At the M step use the pixel margin (summed counts across all detectors) to estimate $\lambda^{(1)}$.
4. Repeat the E step with the new estimate of λ , and iterate to convergence.

For more recent work on speeded EM algorithms for image reconstruction, see Meng and Van Dyk (1997, Section 3.5).

13.4 Loglinear Models for Partially Classified Contingency Tables

13.4.1 The Complete-Data Case

For a complete V -way contingency table with cell probabilities $\{\pi_{jkl\dots u}\}$, it is often desirable to consider more parsimonious models where the cell probabilities have a special structure. For example, independence between the factors corresponds to a model where the probabilities can be expressed in the form

$$\pi_{jkl\dots u} = \tau \tau_j^{(1)} \tau_k^{(2)} \dots \tau_u^{(V)} \quad (13.10)$$

for suitable multiplicative factors τ and $\{\tau_j^{(1)}\}, \{\tau_k^{(2)}\}, \dots, \{\tau_u^{(V)}\}$. It is often helpful to express (13.10) as a loglinear model:

$$\ell n(\pi_{jkl\dots u}) = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \dots + \alpha_u^{(V)}, \quad (13.11)$$

where $\alpha_j^{(1)} = \ell n(\tau_j^{(1)})$, and so forth. Different sets of α 's on the right side of (13.11) yield the same set of cell probabilities $\{\pi_{j_k \dots u}\}$, and V constraints are needed to define the α 's uniquely. A common choice, similar to the choice in analysis of variance, is to set

$$\sum_{j=1}^J \alpha_j^{(1)} = \dots = \sum_{u=1}^U \alpha_u^{(V)} = 0.$$

Equation (13.10) or (13.11) defines a loglinear model for the cell probabilities. A more general class of models is obtained by decomposing the logarithm of the cell probabilities into a sum of constant, main effects as in (13.11), and higher-order associations, and then setting some of the terms in the decomposition to zero. For example, for a $V = 3$ -way table, we write

$$\ell n(\pi_{jkl}) = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \alpha_l^{(3)} + \alpha_{jk}^{(12)} + \alpha_{jl}^{(13)} + \alpha_{kl}^{(23)} + \alpha_{jkl}^{(123)}, \quad (13.12)$$

where the α terms are constrained to sum to zero over any of their subscripts. The terms $\{\alpha_j^{(1)}\}, \{\alpha_k^{(2)}\}, \{\alpha_l^{(3)}\}$ are called the main effects of Y_1, Y_2 , and Y_3 , $\{\alpha_{jk}^{(12)}\}, \{\alpha_{jl}^{(13)}\}, \{\alpha_{kl}^{(23)}\}$ are called two-way associations between Y_1 and Y_2, Y_1 and Y_3 , and Y_2 and Y_3 , respectively, and $\{\alpha_{jkl}^{(123)}\}$ are called three-way associations between Y_1, Y_2 , and Y_3 . Setting all two- and three-way associations to zero yields the independence model (13.11) for $V = 3$ variables. Other models are obtained by setting other terms to zero in (13.12).

An important class of models obtained in this way are *hierarchical* loglinear models, which have the property that inclusion of a V -way association between a set of factors implies inclusion of all $(V - 1)$ -way and lower-order associations and main effects involving subsets of these factors. There are 19 hierarchical models for a three-way table. Nine of them are listed in Table 13.7; the remaining 10 can be obtained by permuting the factors in models (3)–(8) in the table.

An interesting Bayesian alternative to loglinear models for imposing structure on the cell probabilities in multiway contingency tables is proposed in Dunson and Xing (2009). We focus on the loglinear approach here.

ML estimation for hierarchical models varies in complexity depending on the model fitted. In particular, explicit ML estimates can be found for all the models in Table 13.7 except for {12, 23, 31}, where an iterative method, such as iterative proportional fitting (IPF), is necessary.

Two asymptotically equivalent goodness-of-fit statistics are widely used to compare the fit of loglinear models. The likelihood ratio (LR) statistic is

$$G^2 = 2 \sum_c n_c \ell n(n_c / \hat{n}_c), \quad (13.13)$$

Table 13.7 Hierarchical loglinear models for three-way tables

Model	Label	Terms to set to zero in (13.12)
(1)	{123}	None
(2)	{12, 23, 31}	$\{\alpha_{jkl}^{(123)}\}$
(3)	{12, 13}	$\{\alpha_{jkl}^{(123)}, \alpha_{kl}^{(23)}\}$
(4)	{1, 23}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}\}$
(5)	{23}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_j^{(1)}\}$
(6)	{1, 2, 3}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_{kl}^{(23)}\}$
(7)	{2, 3}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_{kl}^{(23)}, \alpha_j^{(1)}\}$
(8)	{1}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_{kl}^{(23)}, \alpha_k^{(2)}, \alpha_l^{(3)}\}$
(9)	{Ø}	$\{\alpha_{jkl}^{(123)}, \alpha_{jk}^{(12)}, \alpha_{jl}^{(13)}, \alpha_{kl}^{(23)}, \alpha_j^{(1)}, \alpha_k^{(2)}, \alpha_l^{(3)}\}$

where the summation is over all the cells c in the table, n_c is the observed count in cell c , and $\hat{n}_c = n\hat{\pi}_c$ is the expected count in cell c estimated from the model. The Pearson chi-squared statistic is defined as

$$X^2 = \sum_c (n_c - \hat{n}_c)^2 / \hat{n}_c. \quad (13.14)$$

If the fitted model is correct, then both G^2 and X^2 are asymptotically chi-squared distributed with degrees of freedom equal to the number of independent restrictions on the cell probabilities. Details on calculating degrees of freedom and more information on loglinear models for complete data are given in Goodman (1970), Haberman (1974), Bishop et al. (1975), and Fienberg (1980). We focus here on the likelihood-ratio statistic (13.13), which tends to perform better in moderate sized samples than (13.14) and is more consistent with our likelihood perspective.

Example 13.8 A Complete Three-Way Table. Table 13.8(a) presents a 2^3 contingency table on survival of infants, previously analyzed in Bishop et al. (1975, Table 1.4-2). Table 13.9 shows estimated cell probabilities and goodness-of-fit statistics for selected loglinear models fitted to these data.

The model {SPC} in Table 13.9(a) places no constraints on the cell probabilities and fits the observed cell proportions perfectly. Hence, the goodness-of-fit statistics are both zero with zero degrees of freedom. Two unsaturated models in Table 13.9(b) and (c) have very low values of G^2 , indicating good fits, namely,

Table 13.8 Example 13.8, a 2^3 contingency table with partially classified observations

Clinic (C)	Prenatal care (P)	Survival (S)		$r = 715$ units
		Died	Survived	
<i>(a) Completely classified cases</i>				
A	Less	3	176	$r = 715$ units
	More	4	293	
B	Less	17	197	$m = 255$ units
	More	2	23	
<i>(b) Partially classified cases (clinic missing)</i>				
	Less	10	150	$m = 255$ units
	More	5	90	

Source: (a) Bishop et al. (1975), table 1.4-2. Reproduced with permission of Springer Nature.

(b) Artificial data.

the model {SC, PC}, which indicates that survival is related to clinic, but survival and prenatal care are not associated conditional on clinic, and the model {SC, PC, SP}, which adds the association SP to the previous model. Because the difference in fits is negligible and the former model is more parsimonious, it will usually be preferred. The model {SP, SC} fits the data poorly and is included for illustrative purposes.

13.4.2 Loglinear Models for Partially Classified Tables

As with the saturated models in Sections 13.2 and 13.3, ML estimation of loglinear models under MAR involves distributing the partially classified counts into the full table using estimated conditional probabilities and then estimating the classification probabilities from the filled-in table. The only difference is that all probabilities are estimated subject to the constraints imposed by the loglinear model. For monotone missingness patterns, these constraints can increase the computational effort required for ML or Bayes, because the factored likelihood method does not apply when the parameters in the factors are not distinct. For nonmonotone patterns, the M step of EM may itself involve iteration, but this is easily addressed by using the ECM algorithm instead of EM, as discussed below.

The standard ML fitting algorithm with complete cross-classified data is IPF, which applies proportional adjustments to the data to successively match margins of the table that are the minimal sufficient statistics under the posited model (e.g., Bishop et al. 1975, chapter 3). The method is described in

Table 13.9 Examples 13.8 and 13.10, estimated cell probabilities $\{\hat{\pi}_{jkl}\} \times 100$ from saturated model {SPC} and three loglinear models, fitted to data in Table 13.8(a)

Clinic (C)	Prenatal care (P)	Survival (S)		Goodness-of-fit
		Died	Survived	
<i>(a) Model: {SPC}</i>				
A	Less	0.42	24.62	$df = 0, G^2 = 0$
	More	0.56	40.98	
<i>(b) Model: {SP, SC, PC}</i>				
A	Less	0.39	24.64	$df = 1, G^2 = 0.04$
	More	0.59	40.95	
B	Less	2.38	27.55	$df = 1, G^2 = 0.04$
	More	0.28	3.22	
<i>(c) Model: {SC, PC}</i>				
A	Less	0.36	24.67	$df = 2, G^2 = 0.08$
	More	0.62	40.92	
B	Less	2.41	27.52	$df = 2, G^2 = 0.08$
	More	0.25	3.24	
<i>(d) Model: {SP, SC}</i>				
A	Less	0.76	35.51	$df = 2, G^2 = 188.1$
	More	0.22	30.08	
B	Less	2.04	16.66	$df = 2, G^2 = 188.1$
	More	0.62	14.11	

Example 8.7 for the case of the simplest loglinear model that does not have an explicit ML solution, the no three-way association model in a $2 \times 2 \times 2$ table. Because IPF increases the likelihood at each iteration, replacing the M step by a single iteration of IPF yields an ECM algorithm (Meng and Rubin 1991), and hence shares similar asymptotic properties with EM.

For Bayesian analyses, the imputation (I) step of DA is unaffected by the model restrictions, and allocates each partially classified count into the set

of possible cells as draws from a multinomial distribution, with conditional probabilities calculated from the previous posterior (P) step. The P step generates a draw from the complete-data posterior distribution of the constrained parameters, with data filled in from the previous I step. In models with explicit complete-data ML estimates, the joint distribution factors into components with unrestricted multinomial parameters. With independent Dirichlet prior distributions for these factors, the posterior distributions are also Dirichlet, and the P step draws from these distributions. In the case of the three-way table described in Table 13.7, this approach applies to all the models except the model with no three-way association {12, 23, 31}.

For models where complete-data ML requires iteration, draws for Bayesian inferences are created using the Gibbs' sampler. The draws of the missing data are as before. Draws of the parameters can be achieved using Bayesian IPF (Gelman et al. 2013), a Bayesian analog of IPF. The CM steps of IPF are replaced by analogous conditional posterior (CP) steps, which are draws of sets of loglinear model parameters of Dirichlet conditional distributions, given current values of the other loglinear model parameters and the imputed data. Specifically, consider the following example, which creates a draw of θ in the limit as $t \rightarrow \infty$:

Example 13.9 Bayesian IPF for the No Three-Way Association Model for a $2 \times 2 \times 2$ Table (Example 8.7 Continued). Suppose we have complete data in a $2 \times 2 \times 2$ table, with $\{y_{ij+}\}$, $\{y_{i+k}\}$, and $\{y_{+jk}\}$ the three two-way margins of the table. At iteration t , let $\{\theta_{ijk}^{(t)}\}$ be current estimates of the cell probabilities, and let $\{y_{ij+}^{(t)}\}$, $\{y_{i+k}^{(t)}\}$, and $\{y_{+jk}^{(t)}\}$ be the imputed two-way marginal counts from the I step. In Bayesian IPF, the CM1 step (8.36) is replaced by the following CP1 step:

$$\text{CP1: } \theta_{ijk}^{(t+1/3)} = \theta_{ij(k)}^{(t)} (g_{ij+}^{(t+1/3)} / g_{+++}^{(t+1/3)}),$$

where $\theta_{ij(k)}^{(t)}$ is the draw at iteration t of the conditional probability defined in Example 8.7, and the imputed proportion $(y_{ij+}^{(t)} / n)$ in (8.36) has been replaced by $(g_{ij+}^{(t+1/3)} / g_{+++}^{(t+1/3)})$, where $\{g_{ij+}^{(t+1/3)}\}$ are draws from the Dirichlet distribution

$$p(\theta_{ij+} | \theta_{i+k}^{(t)}, \theta_{+jk}^{(t)}, Y^{(t)}) \propto \prod_{i=1}^2 \prod_{j=1}^2 \theta_{ij+}^{(\alpha_{ij+} + y_{ij+}^{(t)} - 1)},$$

α_{ij+} are the Dirichlet parameters in the prior distribution of θ_{ij+} , and $g_{+++}^{(t+1/3)} = \sum_{i,j} g_{ij+}^{(t+1/3)}$. Similarly, the CM2 and CM3 steps (8.37) and (8.38) of ECM are replaced by the following CP2 and CP3 steps:

$$\text{CP2: } \theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)} (g_{i+k}^{(t+2/3)} / g_{+++}^{(t+2/3)}),$$

$$\text{CP3: } \theta_{ijk}^{(t+3/3)} = \theta_{ij(k)}^{(t+2/3)} (g_{+jk}^{(t+3/3)} / g_{+++}^{(t+3/3)}),$$

where $\{g_{i+k}^{(t+2/3)}\}$ are draws from the Dirichlet distribution

$$p\left(\theta_{i+k} \mid \theta_{ij+}^{(t)}, \theta_{+jk}^{(t)}, Y^{(t)}\right) \propto \prod_{i=1}^2 \prod_{k=1}^2 \theta_{i+k}^{(\alpha_{i+k} + y_{i+k}^{(t)} - 1)},$$

and $\{g_{+jk}^{(t+3/3)}\}$ are draws from the Dirichlet distribution

$$p\left(\theta_{+jk} \mid \theta_{ij+}^{(t)}, \theta_{i+k}^{(t)}, Y^{(t)}\right) \propto \prod_{j=1}^2 \prod_{k=1}^2 \theta_{ij+}^{(\alpha_{+jk} + y_{+jk}^{(t)} - 1)}.$$

This method extends immediately to any loglinear model. The method first appeared in Gelman et al. (1995, pp. 400–401). An excellent description with examples and a discussion of convergence properties is given by Schafer (1997, pp. 308–320).

Example 13.10 *ML Estimates for an Incomplete Three-Way Table (Example 13.8 Continued).* Suppose that the supplemental data in Table 13.8(b) are appended to the data in Table 13.8(a), analyzed in Example 13.8. Survival (S) and Prenatal Care (P) are observed in the supplemental data, but Clinic (C) is missing. The resulting incomplete data form a monotone pattern with P and S more observed than C.

The likelihood for the combined data in Table 13.8(a) and (b) factors into a factor for the distribution of SP, involving all $r + m = 970$ units, and a factor for the distribution of C given SP, involving the $m = 715$ completely classified units. These two distributions involve distinct parameters for the models {SPC}, {SP, SC, PC}, and {SP, SC}. Hence, ML estimates can be derived for these models by the factored likelihood method of Chapter 7. Table 13.10(a) shows ML estimates of $100\pi_{jkl}$ for the saturated model {SPC}, calculated by the methods of Section 13.2. Table 13.10(b) and (c) show ML estimates for {SP, SC, PC} and {SP, SC}. Because the {SP} margin is fitted in these models, estimates of the probabilities in this margin are the same as those for {SPC}. The conditional probabilities that $C = A$ or B given SP are obtained from the appropriate model applied to the 715 complete units. For {SP, SC}, this calculation is noniterative, but for {SP, SC, PC} it is iterative. The two sets of ML parameter estimates are combined as for the saturated models to obtain ML estimates of the joint probabilities π_{jkl} by Property 6.1 of ML estimates.

Parameters of the distributions of SP and C given SP are not distinct for the model {SC, PC}, so the factored likelihood method cannot be applied to provide ML estimates. Table 13.11 shows four iterations of the EM algorithms for this model; estimates of $100\pi_{jkl}$ are unchanged between iterations 4 and 5, to two decimal places.

Table 13.10 Example 13.10, ML estimates for models {SPC}, {SP, SC, PC}, and {SP, SC} fitted to data in Table 13.8(a) and (b)

Clinic (C)	Prenatal care (P)	Survival (S)	
		Died	Survived
<i>(a) Model: {SPC}</i>			
A	Less	$100(3/20)(30/970) = 0.46$	$100(176/373)(523/970) = 25.44$
	More	$100(4/6)(11/970) = 0.76$	$100(293/316)(406/970) = 38.81$
B	Less	$100(17/20)(30/970) = 2.63$	$100(197/373)(523/970) = 28.49$
	More	$100(2/6)(11/970) = 0.38$	$100(23/316)(406/970) = 3.05$
			Total = 100.0
<i>(b) Model: {SP, SC, PC}</i>			
A	Less	$100(2.8/20)(30/970) = 0.43$	$100(176.2/373)(523/970) = 25.47$
	More	$100(4.2/6)(11/970) = 0.79$	$100(292.8/316)(406/970) = 38.78$
B	Less	$100(17.2/20)(30/970) = 2.66$	$100(196.8/373)(523/970) = 28.45$
	More	$100(1.8/6)(11/970) = 0.34$	$100(23.2/316)(406/970) = 3.07$
			Total = 100.0
<i>(c) Model: {SP, SC}</i>			
A	Less	$100(5.4/20)(30/970) = 0.84$	$100(253.9/373)(523/970) = 36.70$
	More	$100(1.6/6)(11/970) = 0.30$	$100(215.1/316)(406/970) = 28.49$
B	Less	$100(14.6/20)(30/970) = 2.26$	$100(119.1/373)(523/970) = 17.22$
	More	$100(4.4/6)(11/970) = 0.83$	$100(100.9/316)(406/970) = 13.26$
			Total = 100.0

In the preceding example, starting values for the EM algorithm were based on the analysis of the completely classified table. With sparse tables containing zero cells, this procedure can yield unsatisfactory starting values, as discussed in Fuchs (1982). In particular, suppose a marginal table corresponding to a term in the model has an empty cell in the fully categorized table, and the same cell has a positive count in the supplemental table. If starting values are based on the fully categorized table, then the EM algorithm never allows the zero cell to attain a nonzero probability, thus contradicting the supplemental information. This problem can be avoided by forming starting values after adding positive values to the cells of the completely classified table, so that initial estimates are in the interior of the parameter space. In subsequent iterations, these added values can be discarded. Another simpler approach is to create starting values based on an assumption that all variables are mutually independent of each other.

Table 13.11 Example 13.10, ML estimates for model $\{SC, PC\}$ fitted to data in Table 13.8(a) and (b), via the EM algorithm

M step: estimated cell probabilities $\times 100$				E step: filled-in cell counts				Survival	
Iteration	Clinic (C)	Prenatal (P)		Survival (S)		Died		Survived	
		Less	More	Died	Survived	Died	Survived		
1	A	Less	0.36	24.67	3 + (10)(0.36)/2.74 = 4.33	176 + 150(24.62)/52.22 = 246.36	3 + 5(0.62)/0.90 = 7.44	293 + 90(40.92)/44.14 = 376.44	
		More	0.62	40.92	4 + 5(0.62)/0.90 = 7.44				
	B	Less	2.38	27.56	17 + 10(2.38)/2.74 = 25.67	197 + 150(27.56)/52.22 = 276.14			
		More	0.28	3.22	2 + 5(0.28)/0.90 = 3.56			23 + 90(3.22)/44.14 = 29.56	
2	A	Less	0.48	25.42	4.50	246.84			
		More	0.73	38.84	7.56	376.32			
	B	Less	2.72	28.40	25.50	276.16			
		More	0.30	3.12	3.44	29.68			
3	A	Less	0.49	25.42	4.55	246.83			
		More	0.75	38.82	7.59	376.31			
	B	Less	2.69	28.41	25.45	276.17			
		More	0.30	3.12	3.41	29.69			
4	A	Less	0.50	25.42	4.56	246.83			
		More	0.76	38.82	7.60	376.31			
	B	Less	2.68	28.41	25.44	276.17			
		More	0.29	3.12	3.40	29.69			

13.4.3 Goodness-of-Fit Tests for Partially Classified Data

Likelihood-ratio statistics analogous to (13.13) can be calculated for partially classified tables by summing over the cells in the complete and partially classified supplemental tables. Note that unlike the complete-data case, nonzero values of G^2 are obtained for the saturated model ($\{\text{SPC}\}$ in Example 13.10); the values of G^2 for the saturated model provide tests for whether the data are MCAR.

Chi-squared statistics for restricted models can be obtained by calculating G^2 for the restricted model and the saturated model and then subtracting the two quantities (Fuchs 1982). The resulting difference has the same number of degrees of freedom as the LR test for the same model with complete data.

Example 13.11 Goodness-of-Fit Statistics for Incomplete Three-Way Table (Example 13.10 Continued). Goodness-of-fit statistics for the saturated model $\{\text{SPC}\}$ in Example (13.13) are

$$G^2(\text{SPC}) = 7.80, \quad df = 3.$$

To calculate degrees of freedom (df), note that there are $8 + 4 = 12$ cells of data, yielding 11 degrees of freedom for estimating 7 cell probabilities and 1 response probability, or 8 parameters. Hence $df = 11 - 7 - 1 = 3$. Because the 95th percentile of the chi-squared distribution with 3 df is 7.815, the null hypothesis that the data are MCAR yields a p -value of approximately 0.05 for G^2 . The unsaturated models yield

$$G^2(\text{SP, SC, PC}) = 7.84, \quad df = 11 - 6 - 1 = 4,$$

$$G^2(\text{SP, PC}) = 7.84, \quad df = 11 - 5 - 1 = 5,$$

$$G^2(\text{SP, SC}) = 195.92, \quad df = 11 - 5 - 1 = 5.$$

Subtracting the chi-squared values for the saturated model yields

$$\Delta G^2(\text{SP, SC, PC}) = 0.04, \quad \Delta df = 8 - 6 - 1 = 1,$$

$$\Delta G^2(\text{SP, PC}) = 0.20, \quad \Delta df = 8 - 5 - 1 = 2,$$

$$\Delta G^2(\text{SP, SC}) = 188.12, \quad \Delta df = 8 - 5 - 1 = 2,$$

which can be compared with the goodness-of-fit statistics based on the completely classified units in Table 13.9. We conclude, as before, that $\{\text{SP, PC}\}$ is the preferred model.

Problems

- 13.1** Show that for complete data, the Poisson and multinomial models for multiway count data yield the same likelihood-based inferences for the cell probabilities. Show that the result continues to hold when data are MAR.
- 13.2** Derive ML estimates and associated asymptotic sampling variances for the likelihood (13.1). (*Hint:* Remember the constraint that the cell probabilities sum to 1.)
- 13.3** Verify the results of the LR test for the MCAR assumption in Example 13.2.
- 13.4** Compute the fraction of missing information in Example 13.2, using the methods of Section 9.1.
- 13.5** Calculate the expected cell frequencies in the first column of data in Table 13.3(b), and compare the answers with those obtained from complete units.
- 13.6** Suppose that in Example 13.3 there are no units with pattern d . Which parameters are inestimable, in the sense that they do not appear in the likelihood? Estimate the cell probabilities, assuming specific values for the inestimable parameters.
- 13.7** State in words the assumption about the missingness mechanism under which the estimates in Table 13.4(c) are ML for Example 13.4.
- 13.8** Fill in the details in the derivation of Eq. (13.7).
- 13.9** Replicate the calculations of Example 13.4 for estimates of π_{12} .
- 13.10** Redo Example 13.4 assuming that the coarsely classified data in Table 13.4 were summarized as “Improvement” or “No Improvement” (stationary or worse).
- 13.11** Implement the EM algorithm for the data in Table 13.5 with values superscripted a , b and c , d in the supplemental margins interchanged. Compare the ML estimate of the odds ratio $\pi_{11}\pi_{22}\pi_{12}^{-1}\pi_{21}^{-1}$ with the estimate from complete units. Are they identical?

- 13.12** Show that in Example 13.10 the parameters in the factored likelihood are distinct for models {SP, SC, PC} and {SP, SC}, but are not distinct for {SC, PC}.
- 13.13** Display explicit ML estimates for all the models in Table 13.7 except for {12, 23, 31}.
- 13.14** Using results from Problem 13.13, derive the estimates in Table 13.9 for the models {SPC}, {SC, PC}, and {SP, SC}.
- 13.15** Compute ML estimates for the model {SP, SC} for the full data in Table 13.8, with the counts in the supplemental Table 13.8(b) increased by a factor of 10.
- 13.16** Why can starting values including zero probabilities disrupt proper performance of EM? (*Hint:* Consider the loglikelihood.)
- 13.17** Consider bivariate monotone data as in Section 13.2, and suppose the data are MCAR.
- Show that c_j in (13.7) is of smaller order than other terms in the expression.
 - Show that (13.7) is asymptotically equal to

$$\text{Var}(\pi_{jk} - \hat{\pi}_{jk}) \approx \frac{\hat{\pi}_{jk}(1 - \hat{\pi}_{jk})}{r} \left[1 - \frac{\hat{\pi}_{k,j} - \hat{\pi}_{jk}}{1 - \hat{\pi}_{jk}} \frac{n - r}{n} \right].$$

Hence, state the proportionate reduction in asymptotic sampling variance of $\hat{\pi}_{jk}$ over the complete-case estimate, and describe situations when it is large and small. (The analogous situation for normal data is discussed in Section 7.2.1.)

14

Mixed Normal and Nonnormal Data with Missing Values, Ignoring the Missingness Mechanism

14.1 Introduction

In Chapters 11 and 12, we considered a variety of complete-data models for continuous variables, based on the multivariate normal distribution and longer-tailed distributions, with missing data that were missing at random (MAR). The role of categorical variables was confined to that of fully observed covariates in regression models. In Chapter 13, we discussed complete-data models for categorical variables when there were missing values. In this chapter, we consider missing data methods for mixtures of normal and nonnormal variables, with MAR missingness.

Little and Schluchter (1985) discuss a model for missing data with mixed normal and categorical variables and provide relatively simple and computationally feasible expectation–maximization (EM) algorithms with incomplete data. Schafer (1997) discusses Bayes’ inference for this model, and Liu and Rubin (1998) develop a variety of extensions. The basic version of this model is presented in Section 14.2, and extensions are outlined in Section 14.3. Relationships with previously considered algorithms are examined in Section 14.4.

14.2 The General Location Model

14.2.1 The Complete-Data Model and Parameter Estimates

Suppose that the complete data consist of a random sample of size n on K continuous variables (X) and V categorical variables (Y). Categorical variable j has I_j levels so that the categorical variables define a V -way contingency table with $C = \prod_{j=1}^V I_j$ cells. For unit i , let x_i be the $(1 \times K)$ vector of continuous variables and y_i the $(1 \times V)$ vector of categorical variables. Also construct from y_i the

$(1 \times C)$ vector w_i , which equals U_c if unit i belongs to cell c of the contingency table, where U_c is a $(1 \times C)$ vector with 1 as the c th entry and 0s elsewhere.

Let θ denote all unknown parameters. Olkin and Tate (1961) define the “general location model” for the distribution of (x_i, w_i) in terms of the marginal distribution of w_i and the conditional distribution of x_i given w_i :

1. The w_i are independent, identically distributed (iid) multinomial random variables with cell probabilities

$$\Pr(w_i = U_c | \theta) = \pi_c, \quad c = 1, \dots, C; \sum \pi_c = 1. \quad (14.1)$$

2. Given that $w_i = U_c$,

$$(x_i | w_i = U_c, \theta) \sim_{\text{ind}} N_K(\mu_c, \Omega), \quad (14.2)$$

the K -variate normal distribution with mean $\mu_c = (\mu_{c1}, \dots, \mu_{cK})$ and covariance matrix Ω . We write $\Pi = (\pi_1, \dots, \pi_C)$ for the $(1 \times C)$ vector of cell probabilities and $\Gamma = \{\mu_{ck}\}$ for the $(C \times K)$ matrix of cell means of x_i . There are $C - 1 + KC + 1/2K(K + 1)$ parameters, $\theta = (\Pi, \Gamma, \Omega)$, in the model.

The following properties of this model are worth noting:

- In the absence of categorical variables Y , the model reduces to the multivariate normal model in Section 11.2, and the algorithms described here reduce to the corresponding algorithms for multivariate normal data.
- If categorical variables are incomplete and no continuous variables are present, then the data can be arranged as a multiway contingency table with partially classified supplemental margins. The algorithms described here then reduce to maximum likelihood (ML) and Bayes estimation for partially classified contingency tables, as discussed in Chapter 13.
- An important assumption of the basic model is that the within-cell covariance matrix Ω is the same across all the cells of the contingency table. This assumption can be relaxed, as noted in Section 14.5.
- If a particular binary variable (say Y_1), with values 1 and 0, is viewed as a dependent variable, then the conditional probability that $Y_1 = 1$, given the parameters θ and other variables, is $e^L / (1 + e^L)$, where L is linear in the other variables. If Y_1 is the sole categorical variable, then Eqs. (14.1) and (14.2) are the models for two-group discriminant analysis, which is an alternative to logistic regression for predicting Y_1 on the basis of X (see, for example, Press and Wilson 1978).
- If a particular continuous variable (say X_1) is viewed as a dependent variable, then a normal linear regression model results. That is, the conditional distribution of X_1 given the parameters θ and the other variables are normal, with a mean that is a linear combination of the other variables, and constant variance.

Properties (iv) and (v) imply that ML estimates for certain logistic regression models with missing values, and for certain linear regression models with missing continuous and categorical predictors, can be found by finding ML estimates of θ and then transforming them to yield parameters of the appropriate conditional distribution. More details are given in Section 14.4.

The complete-data loglikelihood for this model is

$$\begin{aligned}\ell(\Gamma, \Omega, \Pi) &= \sum_{i=1}^n \ln f(x_i|w_i, \Gamma, \Omega) + \sum_{i=1}^n \ln f(w_i|\Pi) \\ &= h(\Omega) - \frac{1}{2} \text{tr} \left(\Omega^{-1} \sum_{i=1}^n x_i^T x_i \right) + \text{tr} \Omega^{-1} \Gamma \left(\sum_{i=1}^n w_i^T x_i \right) \\ &\quad + \sum_{c=1}^C \left[\left(\sum_{i=1}^n w_{ic} \right) \left(\ln \pi_c - \frac{1}{2} \mu_c^T \Omega^{-1} \mu_c^T \right) \right],\end{aligned}\quad (14.3)$$

where w_{ic} is the c th component of w_i , tr means “trace of the matrix,” and $h(\Omega) = -1/2n\{K\ln(2\pi) + \ln|\Omega|\}$. Maximizing (14.3) yields complete-data ML estimates

$$\begin{aligned}\hat{\Pi} &= n^{-1} \sum w_i, \\ \hat{\Gamma} &= \left(\sum x_i^T w_i \right) \left(\sum w_i^T w_i \right)^{-1}, \\ \hat{\Omega} &= n^{-1} \sum (x_i - w_i \hat{\Gamma})^T (x_i - w_i \hat{\Gamma}),\end{aligned}\quad (14.4)$$

which are simply the observed cell proportions, the observed cell means, and the pooled within-cell covariance matrix of X , respectively.

14.2.2 ML Estimation with Missing Values

Now suppose some of the X s and W s are missing. For unit i , let $x_{(0),i}$ denote the vector of observed continuous variables, $x_{(1),i}$ denote the vector of missing continuous variables, and S_i denote the set of cells in the contingency table where unit i could lie, given the observed categorical variables. We now consider the EM algorithm for ML estimation of θ given data $\{x_{(0),i}, S_i : i = 1, \dots, n\}$.

The density (14.3) belongs to the regular exponential family with complete-data sufficient statistics $\sum x_i^T x_i$, $\sum w_i^T x_i$, and $\sum w_i$, which are, respectively, the raw sum of squares and cross products of the X s, the cell totals of the X s, and the cell counts. Hence, we can apply the simplified form of the EM algorithm of Section 8.4.2. At iteration t the E step computes the expected values of the complete-data sufficient statistics given data $\{x_{(0),i}, S_i : i = 1, \dots, n\}$ and current parameter estimates $\theta^{(t)}$. The contributions from unit i are

E Step:

$$T_{1i}^{(t)} = E \left(x_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right), \quad (14.5)$$

$$T_{2i}^{(t)} = E \left(w_i^T x_i | x_{(0),i}, S_i, \theta^{(t)} \right), \quad (14.6)$$

$$T_{3i}^{(t)} = E \left(w_i | x_{(0),i}, S_i, \theta^{(t)} \right). \quad (14.7)$$

Details of the E step computations are given in Section 14.2.3. The M step computes the complete-data ML estimates (14.4) with complete-data sufficient statistics replaced by their estimates from the E step.

M Step:

$$\begin{aligned} \Pi^{(t+1)} &= n^{-1} \sum_{i=1}^n T_{3i}^{(t)}, \\ \Gamma^{(t+1)} &= H^{-1} \left(\sum_{i=1}^n T_{2i}^{(t)} \right), \\ \Omega^{(t+1)} &= n^{-1} \left[\sum_{i=1}^n T_{1i}^{(t)} - \left(\sum_{i=1}^n T_{2i}^{(t)} \right)^T H^{-1} \left(\sum_{i=1}^n T_{2i}^{(t)} \right) \right], \end{aligned} \quad (14.8)$$

where H is a matrix with elements of $\sum T_{3i}$ along the main diagonal and 0s elsewhere. The algorithm then returns to the E step to recompute (14.5)–(14.7) with the new parameter estimates, and cycles between E and M steps until convergence.

Example 14.1 *ML Analysis of Categorical and Continuous Outcomes in St. Louis Risk Research Data (Example 11.1 Continued).* Little and Schluchter (1985) analyze the data in Table 11.1 from the St. Louis Risk Research Project using the general location model. Recall that there are three categorical variables, risk group of the parent (G), and two outcomes, D_1 = number of symptoms for first child (1 = low, 2 = high) and D_2 = number of symptoms for second child (1 = low, 2 = high). Thus there are $V = 3$ categorical variables that form a $3 \times 2 \times 2$ contingency table with $C = 12$ cells. There are also $K = 4$ continuous variables R_1, V_1, R_2, V_2 , where R_c and V_c are standardized reading and verbal comprehension scores for the c th child in a family, $c = 1, 2$. The variable G is always observed, but the other variables are missing with a variety of different patterns for the units.

ML estimates computed using the EM algorithm under the unrestricted model are displayed in Table 14.1 (Model A). The maximized loglikelihood under the unrestricted model is -872.73 . Perhaps because of the relatively high

Table 14.1 Example 14.1, maximum likelihood estimates for data in Table 11.1

			Cell means											
			Expected frequencies		R_1		R_2		V_1		V_2			
G	D_1	D_2	A	B	A	B	A	B	A	B	A	B	A	B
1	1	1	10.2	4.8	110.2	113.6	99.8	103.0	133.7	140.9	119.4	129.5		
1	1	2	9.0	8.8	123.4	122.8	116.0	115.4	161.1	160.1	132.1	131.0		
1	2	1	3.6	3.7	111.2	105.3	110.0	101.7	147.7	136.9	126.9	111.6		
1	2	2	4.2	9.7	118.0	114.5	111.9	111.1	123.9	120.8	151.4	148.0		
2	1	1	2.2	4.3	87.6	88.4	101.1	101.5	81.1	81.7	103.3	104.2		
2	1	2	7.2	7.8	104.3	104.4	109.4	109.6	134.6	134.8	109.6	109.9		
2	2	1	2.3	3.3	96.4	96.1	134.5	134.3	122.6	122.0	146.1	145.3		
2	2	2	12.3	8.6	106.7	106.6	97.0	96.8	104.3	104.5	102.4	102.3		
3	1	1	2.1	3.2	115.8	115.7	82.9	82.8	137.7	137.5	96.3	96.0		
3	1	2	7.8	5.9	105.7	100.7	100.8	96.1	127.9	119.4	128.3	117.1		
3	2	1	1.0	2.5	56.2	76.2	88.2	108.3	58.3	90.4	105.4	148.6		
3	2	2	7.1	6.4	107.3	107.4	107.0	107.3	107.2	107.2	104.8	104.8		

(b) Standard deviations and correlations														
Standard deviations				Correlations										
Model	R_1	R_2	V_1	V_2	(R_1, R_2)	(R_1, V_1)	(R_1, V_2)	(R_2, V_1)	(R_2, V_2)	(V_1, V_2)	(R_1, R_1)	(R_2, R_2)	(V_1, V_1)	(V_2, V_2)
<i>A</i>	13.2	11.9	20.7	24.1	0.701	0.832	0.825	0.663	0.835	0.885				
<i>B</i>	13.1	11.9	20.1	23.3	0.685	0.832	0.822	0.654	0.836	0.881				

G = risk group of parent, D_c , R_c , V_c = respectively number of symptoms, reading score and verbal score of child c , ($c = 1, 2$), *A*, model with no restrictions on means or cell probabilities; *B*, model with no restrictions on means, cell probabilities restricted so that (D_1, D_2) are jointly independent of G .

degree of missingness for the categorical variables D_1 and D_2 , several local maxima of the loglikelihood were found, and up to 50 iterations were required for convergence of the loglikelihood to two decimal places, depending on the initial estimates used to start EM. Substantial differences were found between a few of the estimated cell means corresponding to different maxima of the loglikelihood (see Little and Schluchter 1985 for details). Such occurrences suggest that drawing inferences requires care, because the data set is not large enough to support conclusions based on assumptions of asymptotic normality.

14.2.3 Details of the E Step Calculations

We now describe in more detail how the quantities $\{T_{1i}^{(t)}, T_{2i}^{(t)}, T_{3i}^{(t)}, i = 1, \dots, n\}$ are computed in Eqs. (14.5)–(14.7). All parameters in the expressions that follow are equal to the current parameter estimates in $\theta^{(t)}$. Calculation of $T_{3i}^{(t)}$ involves finding $E(w_i|x_{(0),i}, S_i, \theta^{(t)})$ for each unit $i = 1, \dots, n$. The c th component of this vector will be denoted as $\omega_{ic} = \Pr(w_i = U_c|x_{(0),i}, S_i, \theta^{(t)})$. That is, for $\alpha^{(0)}, \Omega^{(0)}$, ω_{ic} is the conditional posterior probability that unit i belongs in cell c , given: the observed continuous variables $x_{(0),i}$, the knowledge that unit i is restricted to be in one of the cells in S_i , and $\theta = \theta^{(t)}$. This is positive when $\sum T_{3i}$ where it takes the form

$$\hat{\Omega} = n^{-1} \sum (x_i - z_i \hat{B})^T (x_i - z_i \hat{B}), \quad (14.9)$$

where

$$\delta_{ic} = x_{(0),i} \Omega_{(0),i}^{-1} \mu_{(0),i}^T - \frac{1}{2} \mu_{(0),i} \Omega_{(0),i}^{-1} \mu_{(0),i}^T + \ln(\pi_c) \quad (14.10)$$

and $\mu_{(0),i,c}$ and $\Omega_{(0),i}$ are the mean and covariance matrix in cell c of the continuous variables $x_{(0),i}$ present for unit i .

To calculate $T_{1i}^{(t)}$ and $T_{2i}^{(t)}$, write the continuous variables for unit i as $\{x_{ij}, j = 1, \dots, K\}$. If x_{ij} is missing, define $\hat{x}_{ij}^{(c)} = E(x_{ij}|x_{(0),i}, w_i = U_c, \theta^{(t)})$, the predicted value of x_{ij} from the regression in cell c of X_j on $x_{(0),i}$ evaluated at $\theta = \theta^{(t)}$. The element in the c th row and j th column of $T_{2i}^{(t)}$, for $c = 1, \dots, C$ and $j = 1, \dots, K$, is obtained by multiplying x_{ij} or its estimate by the conditional posterior probability that unit i falls in cell c :

$$E(w_{ic} x_{ij} | x_{(0),i}, S_i, \theta^{(t)}) = \begin{cases} \omega_{ic} \hat{x}_{ij}^{(c)}, & \text{if } x_{ij} \text{ is missing,} \\ \omega_{ic} x_{ij}, & \text{if } x_{ij} \text{ is observed.} \end{cases}$$

When both x_{ij} and x_{ik} are missing, let $\sigma_{jk-(0),i}$ denote the conditional covariance of x_{ij} and x_{ik} given $x_{(0),i}$ and given that $w_i = U_c$. Then the jk th element of $T_{1i}^{(t)}$, for $j, k = 1, \dots, K$, is

$$E(x_{ij} x_{ik} | x_{(0),i}, S_i, \theta^{(t)}) = \sum_{c \in S_i} \omega_{ic} E(x_{ij} x_{ik} | x_{(0),i}, w_i = U_c, \theta^{(t)})$$

$$= \begin{cases} x_{ij} x_{ik}, & x_{ij}, x_{ik} \text{ both observed;} \\ x_{ik} \sum_{c \in S_i} \omega_{ic} \hat{x}_{ij}^{(c)}, & x_{ij} \text{ missing, } x_{ik} \text{ observed;} \\ x_{ij} \sum_{c \in S_i} \omega_{ic} \hat{x}_{ik}^{(c)}, & x_{ik} \text{ missing, } x_{ij} \text{ observed;} \\ \sigma_{jk-(0),i} + \sum_{c \in S_i} \omega_{ic} \hat{x}_{ij}^{(c)} \hat{x}_{ik}^{(c)}, & x_{ij}, x_{ik} \text{ both missing.} \end{cases}$$

where $\sigma_{jk\cdot(0),i}$ is the conditional covariance of X_j and X_k given the set of variables $x_{(0),i}$ observed for unit i . The computations are easily performed by sweep operations, discussed in Section 7.4.3. Consider the matrix

$$Q = \begin{bmatrix} \hat{\Omega}_{(00),i} & \hat{\Omega}_{(01),i}^T & \hat{\Gamma}_{(0),i}^T \\ \hat{\Omega}_{(01),i} & \hat{\Omega}_{(11),i} & \hat{\Gamma}_{(1),i}^T \\ \hat{\Gamma}_{(0),i} & \hat{\Gamma}_{(1),i} & P \end{bmatrix},$$

where P is a $C \times C$ diagonal matrix, having c th diagonal element equal to $2 \ln \pi_c$, for $c = 1, \dots, C$, and $\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{(00),i} & \hat{\Omega}_{(01),i}^T \\ \hat{\Omega}_{(01),i} & \hat{\Omega}_{(11),i} \end{bmatrix}$ and $\hat{\Gamma} = [\hat{\Gamma}_{(0),i} \quad \hat{\Gamma}_{(1),i}]$ are current estimates of Ω and Γ , partitioned according to the observed and missing X variables in unit i . Sweeping on the elements of Q corresponding to observed X s yields

$$\text{SWP}[x_{(0),i}]Q = \begin{bmatrix} G_{11} & G_{12}^T & G_{13}^T \\ G_{12} & G_{22} & G_{23}^T \\ G_{13} & G_{23} & G_{33} \end{bmatrix},$$

where $G_{11} = -\hat{\Omega}_{(00),i}^{-1}$; $G_{12} = \hat{\Omega}_{(00),i}^{-1}\hat{\Omega}_{(01),i}$ are regression coefficients of the missing X s on $x_{(0),i}$; $G_{22} = \hat{\Omega}_{(11),i} - \hat{\Omega}_{(01),i}^T\hat{\Omega}_{(00),i}^{-1}\hat{\Omega}_{(01),i}$ contains the residual variances and $\sigma_{ji\cdot(0),i}$ and covariances $\sigma_{jk\cdot(0),i}$ for $x_{ij}, x_{ik} \in x_{(0),i}$; $G_{13} = \hat{\Omega}_{(00),i}^{-1}\hat{\Gamma}_{(0),i}$ yields the coefficients of $x_{(0),i}$ in the linear discriminant function (14.10); and the c th diagonal element of $\frac{1}{2}G_{33} = \frac{1}{2}P - \frac{1}{2}\hat{\Gamma}_{(0),i}\hat{\Omega}_{(00),i}^{-1}\hat{\Gamma}_{(0),i}^T$ equals the sum of the second and third terms on the right side of (14.10). Thus G_{13} and G_{33} , together with π_c , yield the linear discriminant function δ_{ic} and hence ω_{ic} as in (14.9). Considerable savings in computation is achieved if units with the same pattern of missingness are grouped together to avoid unnecessary sweep operations.

14.2.4 Bayes' Computation for the Unrestricted General Location Model

Draws from the posterior distribution of the parameters of the unrestricted general location model can be obtained by data augmentation (DA, Schafer 1997), with I and P steps paralleling the E and M steps of EM. For simplicity, we assume the noninformative prior distribution for $\theta = (\Pi, \Gamma, \Omega)$:

$$p(\Pi, \Gamma, \Omega) = \prod_{c=1}^C \pi_c^{-1/2} |\Omega|^{-(K+1)/2}.$$

The I step for unit i comprises two substeps, say I1 and I2: I1 imputes the missing categorical variables, which corresponds to assigning the unit to a particular cell of the contingency table formed by the categorical variables. Specifically, unit i is assigned to cell $c \in S_i$ with probability ω_{ic} given by Eq. (14.9), with parameters θ evaluated at current drawn values $\theta^{(t)}$. The I2 step draws values of the missing continuous variables $x_{(1),i}$ from the conditional multivariate normal distribution of $x_{(1),i}$ given $x_{(0),i}$, the cell c determined by the I1 step, and $\theta^{(t)}$. These steps create a completed data set $Y^{(t)}$.

The P step draws values $\theta^{(t+1)}$ of the parameters from their complete-data posterior distribution given $Y^{(t)}$. The new $\Pi^{(t+1)}$ is drawn from the posterior distribution of Π given $Y^{(t)}$, which is Dirichlet with density:

$$p(\Pi | \{Y^{(t)}\}) = \prod_{c=1}^C \pi_c^{n_c^{(t)} - 1/2}, \quad (14.11)$$

where $n_c^{(t)}$ is the number of observed or imputed units in cell c from the previous I1 step. This draw is achieved using the methods of Examples 6.17 and 6.20. The new $\Omega^{(t+1)}$ is drawn from the posterior distribution of Ω given the filled-in data, which is inverse-Wishart:

$$(\Omega | \Pi^{(t+1)}, Y^{(t)}) \sim \text{inv-Wishart}(S^{(t)}, n - C), \quad (14.12)$$

where $S^{(t)}$ is the pooled within-cell covariance matrix of the continuous variables from the filled-in data. The new $\mu_c^{(t+1)}$ ($c = 1, \dots, C$) is drawn from the posterior distribution of μ_c , given $\Omega^{(t+1)}$ and $Y^{(t)}$, which is multivariate normal:

$$\left(\mu_c | \Pi^{(t+1)}, \Omega^{(t+1)}, Y^{(t)} \right) \sim N_K \left(\bar{y}_c^{(t)}, \Omega^{(t+1)} / n_c^{(t)} \right), \quad (14.13)$$

where $\bar{y}_c^{(t)}$ is the mean of the filled-in continuous variables in cell c . The draws in Eqs. (14.12) and (14.13) are carried out using the methods of Examples 6.18 and 6.21.

Example 14.2 Bayes' Analysis of St. Louis Data (Example 14.1 Continued). The DA algorithm for the unrestricted general location model was applied to the St. Louis data in Example 14.1. Posterior means and posterior standard deviations of the cell probabilities and cell means are shown in Table 14.2, which can be compared with the ML estimates in Table 14.1. Figures 14.1A displays sequences of 10 000 successive draws, and histograms of the last 8000 draws, for the means of the four outcomes in cell (1,1,1). Figure 14.1B shows analogous results for four transformed covariance parameters. Note that the posterior means of the cell means are quite different from the ML estimates for some cells and have large associated posterior standard deviations. These results reflect the sparse data and relatively flat likelihoods. The DA or Gibbs'

Table 14.2 Example 14.1, posterior means and standard deviations of parameters from data augmentation applied to the data in Table 11.1, unrestricted general location model

			Cell means											
			Expected frequencies		R_1		R_2		V_1		V_2			
G	D_1	D_2	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	1	11.8	3.9	112.0	5.9	101.6	4.8	136.6	10.8	119.8	9.2		
1	1	2	7.9	2.9	122.2	7.2	117.7	6.5	155.7	11.6	132.8	11.6		
1	2	1	3.0	1.8	109.5	10.5	99.3	12.8	132.4	17.4	143.6	18.3		
1	2	2	3.9	2.1	124.9	13.9	118.8	9.4	144.0	17.5	142.2	18.1		
2	1	1	1.8	1.2	93.9	13.1	95.3	13.4	90.1	23.7	91.6	24.7		
2	1	2	6.0	2.5	101.6	7.9	109.0	6.7	132.9	18.5	113.5	16.4		
2	2	1	2.2	1.4	91.9	12.5	86.1	21.1	74.6	33.4	101.3	31.3		
2	2	2	14.0	3.5	104.8	5.1	103.2	4.5	109.8	9.2	108.8	8.3		
3	1	1	3.0	1.6	103.7	10.3	86.0	9.0	122.0	16.2	101.8	17.8		
3	1	2	6.2	2.4	120.4	11.9	100.1	6.8	132.4	18.3	120.6	16.3		
3	2	1	1.9	1.2	82.8	14.0	89.9	12.6	101.1	23.1	109.3	24.9		
3	2	2	7.2	2.6	104.8	7.1	107.5	6.0	104.1	13.5	113.8	14.5		

G = risk group of parent, D_c , R_c , V_c = respectively number of symptoms, reading score and verbal score of child c , ($c = 1, 2$).

sampler sequences and histograms of draws look reasonably stable for the means, but the sequences for the covariances display some “jumpiness,” reflecting lack of information to estimate some of these parameters. We prefer the Bayesian results to ML, because they tend to average over plausible regions of the likelihood and better reflect the variability in the data. Bootstrap standard errors for the ML estimates (not shown here) are generally somewhat smaller than the posterior standard deviations and are less reflective of the true variability in this sparse dataset.

14.3 The General Location Model with Parameter Constraints

14.3.1 Introduction

The model of Section 14.2 specifies a distinct mean vector μ_c for each cell c of the table and makes no restrictions on the cell probabilities, other than the obvious restriction that $\sum \pi_c = 1$. In this section, we describe a model that puts

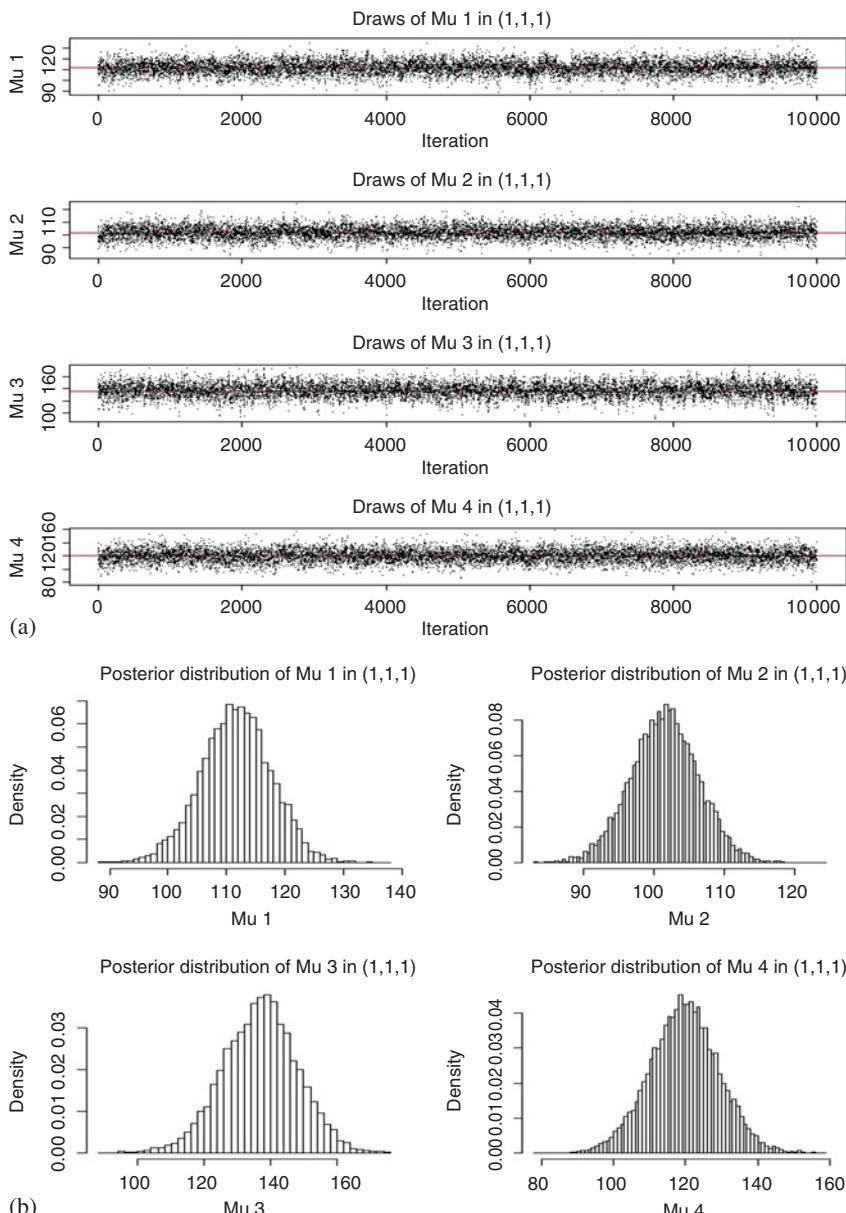


Figure 14.1A Example 14.2, (a) Sequences of draws from the posterior distributions of the means (1, R_1 ; 2, R_2 ; 3, V_1 ; 4, V_2) in cell (1,1,1). (b) Histograms of the posterior distributions for each mean.

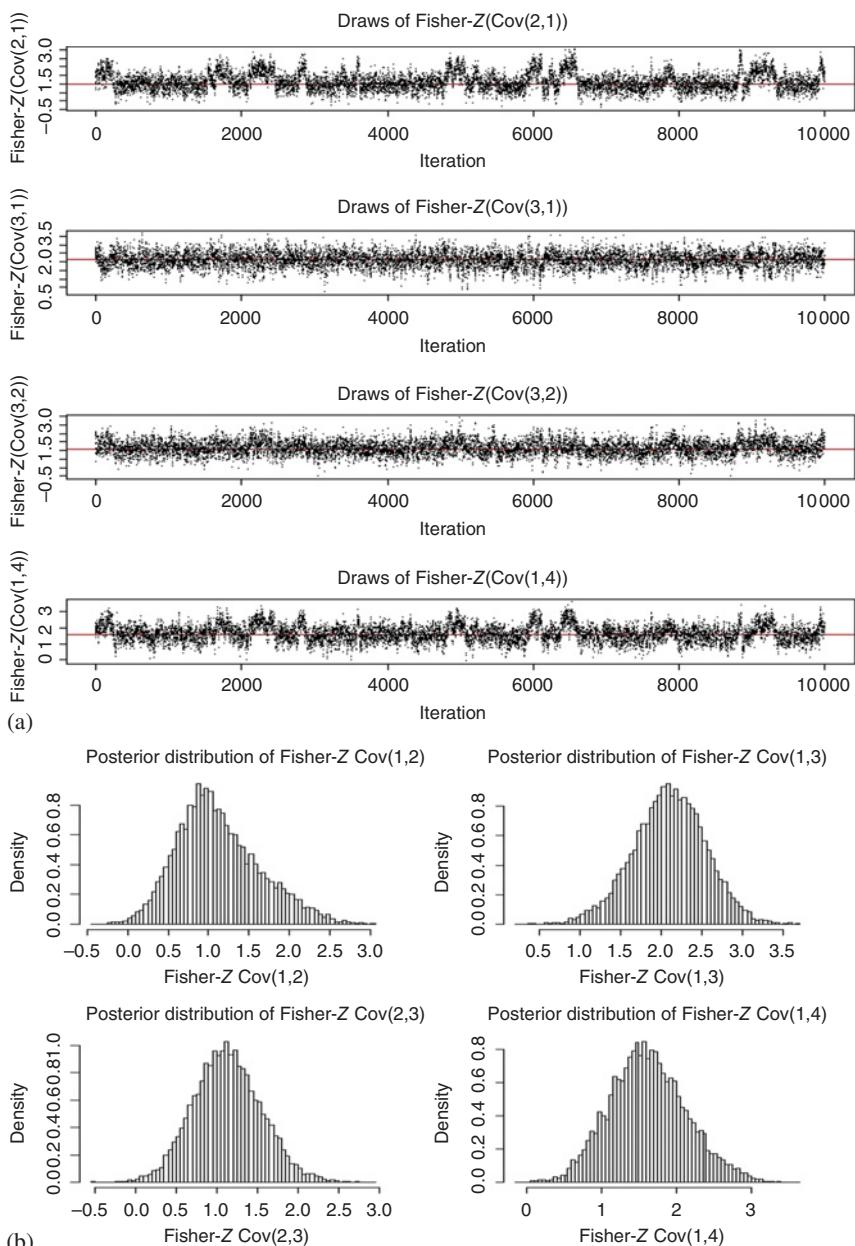


Figure 14.1B Example 14.2, (a) Sequences of Gibbs' sampler draws from the Fisher-Z transformed posterior distributions of selected covariances. (b) Histograms of the posterior distributions for each parameter.

analysis of variance (ANOVA)-like restrictions on the $\{\mu_c\}$, and models the $\{\pi_c\}$ by a restricted loglinear model. This more complex model is considered for complete-data discriminant analysis by Krzanowski (1980, 1982).

14.3.2 Restricted Models for the Cell Means

For $u \leq C$, let z_i be a $1 \times u$ vector of design variables for unit i , which can be obtained from the cell indicator vector w_i as $z_i = w_i A$, where A is a known $C \times u$ matrix that represents the chosen model. The more general model specifies that the conditional distribution of x_i given w_i depends on w_i only through z_i , in that $f(x_i | w_i, \theta) \sim N_k(z_i B, \Omega)$, where B is a $(u \times k)$ matrix of unknown parameters, $\theta = (B, \Omega, \Gamma)$. Note that $E(x_i | w_i, \theta) = w_i AB$, so that $\Gamma = AB$. In the model of Section 14.2, A is the $C \times C$ identity matrix.

14.3.3 Loglinear Models for the Cell Probabilities

Another way of reducing the number of model parameters is to constrain the cell probabilities Π by a loglinear model, as discussed in Section 13.4. For example, suppose the cells are formed by a joint classification of $V = 3$ categorical variables Y_1 , Y_2 , and Y_3 with, respectively, I_1 , I_2 , and I_3 levels, and $C = I_1 \times I_2 \times I_3$. We modify the notation so that π_{jkl} is the probability that $Y_1 = j$, $Y_2 = k$, $Y_3 = l$, for $j = 1, \dots, I_1$, $k = 1, \dots, I_2$, and $l = 1, \dots, I_3$. The loglinear models are obtained by writing

$$\ln \pi_{jkl} = \alpha + \alpha_j^{(1)} + \alpha_k^{(2)} + \alpha_l^{(3)} + \alpha_{jk}^{(12)} + \alpha_{jl}^{(13)} + \alpha_{kl}^{(23)} + \alpha_{jkl}^{(123)},$$

and setting subsets of the α terms equal to 0 (see Section 13.4 for more details).

14.3.4 Modifications to the Algorithms of Previous Sections to Accommodate Parameter Restrictions

For a general V -way table with $C = \prod_{j=1}^V I_j$ cells, let α denote the nonzero α terms in the loglinear model, and write $\pi_c(\alpha)$ for the constrained probability of a unit falling in cell c , $c = 1, \dots, C$. We now sketch modifications of the algorithms in Sections 14.2.2 and 14.2.3 when the reduced models in Sections 14.3.2 and 14.3.3 are fitted to incomplete data.

For a particular choice of the models in Sections 14.3.2 or 14.3.3, let $\alpha^{(0)}$, $\Omega^{(0)}$, and $B^{(0)}$ be initial estimates of the parameters, perhaps calculated from a starting assumption that all variables are independent. Also let $\Gamma^{(0)} = AB^{(0)}$, where A is a known matrix, and $\pi_c^{(0)} = \pi_c(\alpha^{(0)})$, $c = 1, \dots, C$. The restricted models of Sections 14.3.2 and 14.3.3 are regular exponential family models, with complete-data minimal sufficient statistics $\sum x_i^T x_i$, $\sum w_i^T w_i A$ and linear combinations of the counts $\sum w_i$ determined by the margins fitted in the log-linear model. Because these quantities are linear functions of the complete-data

sufficient statistics for the model in Section 14.2, the E step for iteration t computes $\sum T_{1i}^{(t)}$, $\sum T_{2i}^{(t)}$, and $\sum T_{3i}^{(t)}$ via Eqs. (14.5)–(14.7), and then computes the linear combinations of these functions that yield the complete-data minimal sufficient statistics for the reduced model. For Bayesian computations, the I step is the same as for the unrestricted model.

The M step and P step calculations differ from those for the unrestricted model, yielding estimates of Γ , Ω , and Π that satisfy the model restrictions. For estimates of the loglinear model parameters α , first form the multiway table with cell frequencies given in the vector $\sum T_{3i}^{(t)}$ (Eq. (14.7)), which includes fractional entries from the partially classified counts distributed into the table in the E step. The updated estimates of α are obtained by fitting the assumed loglinear model to the counts in $\sum T_{3i}^{(t)}$ by a complete-data method. If explicit estimates are not available, one step of IPF can be taken to update the estimate of α , turning this EM algorithm into an ECM algorithm. For Bayesian computations, Bayesian IPF, as discussed in Chapter 13, can be used to create updated draws of the parameters. The probabilities in the fitted table are the new estimates of $\{\pi_c(\alpha)\}$, used for the next M step.

With complete data, the ML estimates of B and Ω (e.g., see Anderson 1965, Chapter 8) are $\hat{B} = (\sum z_i^T z_i)^{-1} \sum z_i^T x_i$ and $\hat{\Omega} = n^{-1} \sum (x_i - z_i \hat{B})^T (x_i - z_i \hat{B})$. The M step estimates of B and Ω are obtained by setting $z_i = w_i A$ in the preceding equations for \hat{B} and $\hat{\Omega}$, and replacing $\sum x_i^T x_i$, $\sum w_i^T x_i$, and $\sum w_i^T w_i$ by $\sum T_{1i}^{(t)}$, $\sum T_{2i}^{(t)}$, and $D^{(t)}$, respectively, where $D^{(t)}$ is a matrix with elements of $\sum T_{3i}^{(t)}$ on the diagonal and zeros elsewhere. The updated estimates of B , Γ , and Ω in the M step of iteration t are then

$$B^{(t+1)} = (A^T D^{(t)} A)^{-1} A^T \left(\sum T_{2i}^{(t)} \right), \quad (14.14)$$

$$\Gamma^{(t+1)} = AB^{(t+1)}, \quad (14.15)$$

and

$$\Omega^{(t+1)} = n^{-1} \left[\sum_{i=1}^n T_{1i}^{(t)} - \left(\sum_{i=1}^n T_{2i}^{(t)} \right)^T A (A^T D^{(t)} A)^{-1} A^T \left(\sum_{i=1}^n T_{2i}^{(t)} \right) \right]. \quad (14.16)$$

When no restrictions are placed on the means, A is the $C \times C$ identity matrix and the equations for $\Omega^{(t+1)}$ and $\Gamma^{(t+1)}$ in (14.14)–(14.16) are equivalent to their counterparts in Eq. (14.8). The new estimates $\Pi^{(t+1)}$, $\Gamma^{(t+1)}$, and $\Omega^{(t+1)}$ are then input to the next E step, given by Eqs. (14.5)–(14.7).

The P step for Bayesian computations first draws $\Omega^{(t+1)}$ from an inverse-Wishart distribution, as for the unrestricted unit given by Eq. (14.12), but with $S^{(t)}$ replaced by the right side of (14.16) and the degrees of freedom $n - C$ replaced by $n - u$. Then $B^{(t+1)}$ is drawn from a multivariate normal distribution centered at the right side of (14.14), with covariance matrix $\Omega^{(t+1)}$.

Example 14.3 Restricted Models for St. Louis Data (Example 14.1 Continued). In Section 14.2.2, the unrestricted location model was fitted to the data in Table 11.1. The model has too many parameters for a reasonable analysis, with 69 parameters for 69 incomplete units. In this section, we fit and test models with fewer parameters that correspond to hypotheses of substantive interest. In particular, suppose we wish to assess the hypothesis that the occurrence of adverse psychiatric symptoms in children is unrelated to the risk group of the parent, which implies that

$$\pi_{jkl} = \pi_{j++}\pi_{+kl}, \quad j = 1, 2, 3; \quad k, l = 1, 2,$$

where π_{jkl} is the probability associated with level j of G , and levels k and l , respectively, of D_1 and D_2 . No restrictions are placed on the cell means of the continuous variables. Little and Schluchter (1985) fit this constrained model to the data, using the method of Section 14.3.4.

ML estimates for the restricted model are shown in Table 14.1 (Model B). The maximized loglikelihood was -877.64 . Recall that the loglikelihood for the full model fitted in Section 14.2.2 was -872.73 . The likelihood ratio (LR) statistic for testing whether D_1 and D_2 are independent of G was thus $2(-872.73 + 877.64) = 9.82$ with 6 degrees of freedom, suggesting little evidence of lack of fit. Another local maximum (loglikelihood -877.72) was found for this model.

In search for simpler models, Little and Schluchter (1985) next fit the model where the $G \times D_1$, $G \times D_2$, and $G \times D_1 \times D_2$ interaction effects on the means of the continuous variables were set to 0, with the same restricted model for the cell probabilities. The restrictions on the means of continuous variables can be written as $E(x_i | z_i, \theta) = z_i B$, where B is a 6×4 matrix of parameters, and $z_i = w_i A$, where

$$A^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{bmatrix},$$

and the 12 cells in the vector w_i are arranged such that the index of D_2 changes fastest and the index of G changes most slowly. This model reduces the number of parameters needed to describe the means from 48 to 24.

Again, multiple local maxima of the likelihood function were found. In spite of this, Frumento et al. (2016) suggest that comparing scaled LR statistics of the form

$$\text{SLR} = 2(\ln L_1 - \ln L_2)/\text{df},$$

Units	Variables					
	Y_1	...	Y_V	X_1	...	X_K
1	0	...	0	×	...	×
:	:		:	:		:
r	0	...	0	×	...	×
$r+1$	×	...	×	1	...	1
:	:		:	:		:
n	×	...	×	1	...	1

Figure 14.2 Pattern of missing data leading to simple ML estimates. 0, Observed; 1, missing; and \times , observed or missing. Source: Little and Schluchter (1985). Reproduced with permission of Oxford University Press.

where L_1 and L_2 are the maximized likelihoods under the two models and df is the difference in the number of parameters, can be informative. The largest maximized loglikelihood for this model was found to be -910.46 , leading to a scaled likelihood ratio compared to the full model of $SLR = 2(910.46 - 873.73)/30 = 2.45$, suggesting that the reduced model does not fit the data. The authors also fit the model where only the three-way $G \times D_1 \times D_2$ interaction effect was set to 0, with the same restriction on cell probabilities. This model gave a scaled likelihood ratio compared to the full model of $59.39/14 = 4.24$, again evidence that the reduced model does not fit the data. These results suggest that the degree to which parental mental health affects reading and verbal comprehension performance in the child depends on the psychological state of the child, as one would expect.

14.3.5 Simplifications When Categorical Variables are More Observed than Continuous Variables

The algorithms of Sections 14.2 and 14.3 simplify for the data pattern of Figure 14.2, where the V categorical variables are more observed than the K continuous variables. That is, all the categorical variables are observed for units where one or more of the continuous variables are observed. The incomplete-data likelihood then factors into the likelihood for the marginal distribution of (Y_1, \dots, Y_V) and the likelihood for the conditional distribution of (X_1, \dots, X_K) given (Y_1, \dots, Y_V) . ML estimates for the model of Section 14.3 can be obtained as follows:

1. Estimate the parameters of the joint distribution of Y from the first V columns of Figure 14.2. Because these data are entirely categorical, ML algorithms for partially classified contingency tables apply here.
2. Estimate the parameters of the conditional distribution of X given Y from the first r rows of Figure 14.2. The multivariate normal EM algorithm can be used here, even though categorical variables are present. Dummy

variables representing the effects z_i in the ANOVA design are included in the multivariate normal EM algorithm, treating them as if they were continuous variables. Elements corresponding to these variables are then swept in the final estimated covariance matrix of all the variables, yielding estimates $\hat{\beta}, \hat{\Omega}$ of the parameters of the conditional distribution of X given Y . These are ML by an application of the theory of Chapter 7 for factored likelihoods.

Analogous simplifications occur for Bayes' estimation; details are omitted.

14.4 Regression Problems Involving Mixtures of Continuous and Categorical Variables

14.4.1 Normal Linear Regression with Missing Continuous or Categorical Covariates

The methods of Sections 14.2 and 14.3 can be readily applied to yield algorithms for linear regression with missing data. It is easily shown that the general location model (14.1) and (14.2) implies that the conditional distribution of a continuous variable (say X_1) given the other variables is

$$(X_1 | X_2, \dots, X_K, Y_1, \dots, Y_V, \theta) \sim N \left(\beta_{c0}(\theta) + \sum_{j=2}^K \beta_j(\theta) X_j, \sigma^2(\theta) \right), \quad (14.17)$$

where c is the cell of the contingency table formed by (Y_1, \dots, Y_V) , and the parameters are written as functions of the general location model parameters θ . By Property 6.1 or 6.1B, ML or Bayes for the general location model also yield ML or Bayes inference for the model (14.17) with missing data on the outcome and/or the regressors, by computing the regression parameters $\{\beta_{c0}(\theta), \beta_j(\theta), \sigma^2(\theta)\}$ with θ replaced by ML estimates $\hat{\theta}$ or draws $\theta^{(d)}$ from the posterior distribution. The MANOVA restrictions for the distribution of X given Y , discussed in Section 14.3, induce corresponding restrictions on the coefficients $\{\beta_{c0}(\theta)\}$ of (14.17), yielding other regression models.

For the special case of a single continuous outcome X and regressors that are all categorical, the M step of EM is weighted linear regression, with weights based on current estimates (14.9) of the probabilities that an incomplete unit i belongs to each of the set of possible cells S_i . This idea extends readily to the more general class of generalized linear models (Example 6.11) with incomplete categorical covariates. Ibrahim (1990) calls the resulting EM algorithm the method of weights. For more recent developments, see Horton and Laird (1998) and Ibrahim et al. (1999). In a similar vein, Schluchter and Jackson (1989) discuss EM for survival analysis with incomplete categorical covariates.

When the continuous variables are completely observed and Y comprises a k -category covariate that is entirely missing, then the general location model,

(14.1) and (14.2), reduces to that of Day (1969) for k multivariate normal mixtures, which provides a parametric form of cluster analysis. Because the algorithm still works with incompletely recorded continuous variables, it provides an extension of Day's algorithm to incomplete data. As with many mixture models, multiple maxima of the likelihood are a definite possibility (Aitkin and Rubin 1985), so it is advisable to apply the algorithm for a variety of choices of starting values for the parameters.

Example 14.4 A Univariate Mixture Model for Biological Data. Aitkin and Wilson (1980) examined the behavior of the EM algorithm for mixture models on several small data sets. One example was Darwin's data on differences in heights of pairs of self-fertilized and cross-fertilized plants, displayed in Table 14.3(a). The standard ML estimates, assuming a single normal sample with mean μ and variance σ^2 , are displayed in Table 14.3(b) along with the value of $-2 \log\text{likelihood}$ (omitting the constant $n \ln(2\pi)$). A two-component normal mixture with means μ_1 and μ_2 , common variance σ^2 , and mixing proportion p was fit to these data, using EM starting from a variety of initial values. All starting values were obtained by specifying an initial guess as to which

Table 14.3 Example 14.4, results of the EM for mixtures applied to Darwin's data

(a) Darwin's data on differences in heights of self-fertilized and cross-fertilized plants														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
-67	-48	6	8	14	16	23	24	28	29	41	49	56	60	75
(b) Results assuming one normal population														
-2 loglikelihood					$\hat{\mu}$					$\hat{\sigma}^2$				
122.9					20.93					1329.7				
(c) Results of EM iterations for two-component normal with common variance														
Starting specifications (units in component 1)			-2 loglikelihood			$\hat{\mu}_1$		$\hat{\mu}_2$		$\hat{\sigma}^2$		$15\hat{\mu}$		
1			116.0			-57.4		33.0		385.4		2.00		
15			122.9			21.62		20.91		1330.0		0.957		
Any subset of {1,..., 9}			116.0			-57.4		33.0		385.4		2.00		
Any subset of {10,..., 15}			122.9			Estimates vary, depending on exact starting value								

units belonged to component 1 and component 2 (i.e., initial probabilities of component membership were all zero or one), and then applying the M step to obtain initial parameter estimates. The results of these iterations are displayed in Table 14.3(c) and demonstrate the sensitivity of the final estimate to starting values. The likelihood function appears to be bimodal, with a peaked mode at the estimates obtained starting from the first or third starting values and a lower broad mode at the estimates obtained from the second starting value.

14.4.2 Logistic Regression with Missing Continuous or Categorical Covariates

Now suppose that a binary variable, say Y_1 , is the dependent variable. The general location model implies that the conditional distribution of Y_1 given (Y_2, \dots, Y_V) and (X_1, \dots, X_K) is Bernoulli, with

$$\text{logit}(\Pr(Y_1 = 1 | Y_2, \dots, Y_V, X_1, \dots, X_K), \theta) = \gamma_{d0}(\theta) + \sum_{j=1}^K \gamma_{dj}(\theta)X_j, \quad (14.18)$$

where d indexes the cell defined by the values of (Y_2, \dots, Y_V) , and θ again represents the location model parameters. ML or Bayes inference for the model (14.18) with incomplete data is obtained by fitting the general location model, and computing the regression parameters $\{\gamma_{d0}(\theta), \gamma_{dj}(\theta)\}$ in (14.18) with θ replaced by ML estimates $\hat{\theta}$ or draws $\theta^{(d)}$ from its posterior distribution. Restrictions on the parameters θ in (14.18) yield other logistic models. For more discussion of various methods for analyzing incomplete data in logistic regression (see Vach 1994).

As discussed in Chapter 10, an alternative way to implement Bayesian inference is to multiply-impute the missing values based on the model (14.1) and (14.2), and then combine complete-data inferences using the multiple-imputation (MI) methods in Section 10.2. An interesting alternative approach is to multiply-impute using the general location model as before, but modify the complete-data analysis, as follows: instead of fitting the general location model and transforming the parameters, estimate the parameters of (14.8) directly by a standard logistic regression applied to each filled-in data set, that is, apply the model

$$(Y_1 | Y_2, \dots, Y_V, X_1, \dots, X_K) \sim \text{Bern}\left(\gamma_{d0} + \sum_{j=1}^K \gamma_{dj}X_j\right).$$

An advantage of this approach is that the normality assumptions (14.2) of the general location model are only used to fill in the missing values and are not required for the complete-data analysis, which fixes the covariates. As a result, MI inferences are less sensitive to normality assumptions than ML or Bayes

based on the general location model for the joint distribution, particularly when the fraction of missing information is small, so little is imputed.

14.5 Further Extensions of the General Location Model

The general location model has categorical variables marginally distributed as multinomial and continuous variables conditionally normally distributed with different means across cells defined by the categorical variables, but a common covariance matrix across cells. Two extensions of the model are obtained by (i) replacing the common covariance matrix across cells with different but proportional covariance matrices, where the proportionality constants are to be estimated; and (ii) replacing the multivariate normal distributions in the model with multivariate t distributions, where the degrees of freedom can also vary across cells and are to be estimated. The t distribution is just one example of more general ellipsoidally symmetric distributions that can be used in place of the normal. These extensions can provide more accurate fits to real data and can be viewed as tools for robust inference. Moreover, the models can be useful for multiple imputation of missing values, assuming an ignorable missingness mechanism. Liu and Rubin (1998) discuss ML estimation for these extensions using the alternating expectation conditional maximization (AECM) algorithm of Section 8.5.2. They also present a monotone-data DA scheme for drawing parameters and missing values from their joint posterior distribution.

Problems

- 14.1 Show that Eq. (14.4) provides ML estimates of the parameters for the complete-data loglikelihood Eq. (14.3).
- 14.2 Using the factored likelihood methods of Chapter 7, derive ML estimates of the general location model for the special case of one fully observed categorical variable Y and one continuous variable X with some missing values.
- 14.3 Suppose that in Problem 14.2, X is fully observed and Y has some missing values. Show that ML estimates for the general location model cannot be found by factoring the likelihood, because the parameters of the appropriate factorization are not distinct. Suggest an alternative model for which the parameters of the factorization are distinct and display ML estimates for this model.

- 14.4** Compare the properties of discriminant analysis and logistic regression for classifying units into groups on the basis of known covariates (see, for example Press and Wilson 1978; Krzanowski 1980, 1982).
- 14.5** Using Bayes' theorem, show that Eq. (14.9) follows from the definition of the general location model, Eqs. (14.1) and (14.2).
- 14.6** Derive the expressions in Section 14.2.3 for the conditional expectations of $w_{im}x_{ij}$ and $x_{ij}x_{ik}$ given $x_{(0),i}$, S_i and $\theta^{(t)}$, from properties of the general location model.
- 14.7** A survey of 20 graduates of a university class five years after graduation yielded the following results for the variables sex (1 = male, 2 = female), race (1 = white, 2 = other), and annual income, measured on a log scale (? denotes missing):

Unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Sex	1	1	1	2	2	2	2	2	2	2	1	1	2	2	1	1	1	2	2	
Race	1	1	1	1	1	1	1	1	1	1	2	2	2	2	?	?	?	?	?	
Income	25	46	31	5	16	26	8	10	2	?	?	20	29	?	32	?	?	38	15	

- (a) Compute ML estimates for the general location model applied to these data, based on complete units only.
- (b) Develop explicit formulas for the E and M steps (14.5)–(14.8) for these data, and carry out three steps of the EM algorithm, starting from estimates found in (a).
- 14.8** Repeat (b) of Problem 14.7, with the restriction that the variables race and sex are independent.
- 14.9** Derive the maximized loglikelihood of the data in Problem 14.7 for the models of Problems 14.7 and 14.8, and hence derive the likelihood ratio statistic for testing independence of race and sex. Note that the sample size is too small for this statistic to be considered distributed as chi-squared for this illustrative data set (for help, see Little and Schluchter 1985).
- 14.10** Describe the Bayesian analog of the simplified ML algorithm in Section 14.3.5.

- 14.11** Derive Eq. (14.17) from Eqs. (14.1) and (14.2), and hence express the parameters $\{\beta_{c0}, \beta_j, \sigma^2\}$ as functions of the general location model parameters $\theta = (\Pi, \Gamma, \Omega)$. Consider the impact of imposing constraints on θ , the parameters of the linear model (14.17).
- 14.12** Derive Eq. (14.18) from Eqs. (14.1) and (14.2), and hence express the parameters $\{\gamma_{d0}, \gamma_{dj}\}$ as functions of the general location model parameters $\theta = (\Pi, \Gamma, \Omega)$. Consider the impact of imposing constraints on θ and on the parameters of the logistic model (14.18).

15

Missing Not at Random Models

15.1 Introduction

The examples and methods in Chapters 7–14 were based on the ignorable likelihood:

$$L_{\text{ign}}(\theta|Y_{(0)}, X) \propto f(Y_{(0)}|X, \theta), \quad (15.1)$$

regarded as a function of the parameter θ for fixed observed data $Y_{(0)}$; in (15.1), X represents fully observed covariates, and $f(Y_{(0)}|X, \theta)$ is obtained by integrating the missing data $Y_{(1)}$ out of the density $f(Y|X, \theta) = f(Y_{(0)}, Y_{(1)}|X, \theta)$. In Chapter 6, we showed that sufficient conditions for basing inference about θ on (15.1), rather than the full likelihood from a model for Y and M given X , are that (i) the missing data are missing at random (MAR) and (ii) the parameters θ and ψ are distinct, as defined in Section 6.2. In this chapter, we consider situations where the missingness mechanism is missing not at random (MNAR), and valid ML, Bayesian, and multiple-imputation (MI) inferences generally need to be based on the full likelihood:

$$L_{\text{full}}(\theta, \psi|Y_{(0)}, X, M) \propto f(Y_{(0)}, M|X, \theta, \psi), \quad (15.2)$$

regarded as a function of the parameters θ, ψ for fixed observed data $Y_{(0)}$ and missingness pattern M ; here $f(Y_{(0)}, M|X, \theta, \psi)$ is obtained by integrating $Y_{(1)}$ out of the joint density $f(Y, M|X, \theta, \psi)$ based on a joint model for Y and M given X .

Two main approaches for formulating MNAR models can be distinguished. We consider them for situations where the units' values of M and Y are modeled as independent given X , that is, $f(M, Y|X, \theta, \psi) = \prod_{i=1}^n f(m_i, y_i|x_i, \theta, \psi)$ – these models may not be independent, identically distributed (iid) because of the conditioning on x_i , which generally varies with i . *Selection* models factor the joint distribution of m_i and y_i as

$$f(m_i, y_i|x_i, \theta, \psi) = f(y_i|x_i, \theta)f(m_i|x_i, y_i, \psi), \quad (15.3)$$

where densities are distinguished by their arguments. The first factor characterizes the distribution of y_i in the population, the second factor models the missingness mechanism, and θ and ψ are distinct. This factorization underlies the theory of Section 6.2. Alternatively, *pattern-mixture* models factor the joint distribution as

$$f(m_i, y_i | x_i, \xi) = f(y_i | x_i, m_i, \xi) f(m_i | x_i), \quad (15.4)$$

where the first distribution characterizes the distribution of y_i , given x_i in the strata defined by different patterns of missingness, m_i ; the second distribution models the probabilities of the different patterns (Rubin 1977; Glynn et al. 1986, 1993; Little 1993c), and ξ are distinct. The essential distinction between the two factorizations (15.3) and (15.4) becomes clear when considering specific examples, which we now do.

Example 15.1 Pattern-Mixture and Selection Models for Univariate Nonresponse. Suppose for simplicity that missing values are confined to a single variable. Let $y_i = (y_{i1}, y_{i2})$, where y_{i1} is fully observed, and scalar y_{i2} is observed for $i = 1, \dots, r$ but missing for $i = r + 1, \dots, n$. Let m_{i2} be the missingness indicator for y_{i2} , with $m_{i2} = 1$ if y_{i2} is missing and $m_{i2} = 0$ if y_{i2} is observed. A pattern-mixture model factors the density of $Y_{(0)}$ and M given X as

$$\begin{aligned} f(Y_{(0)}, M | X, \xi) &= \prod_{i=1}^r f(y_{i1}, y_{i2} | x_i, m_{i2} = 0, \xi) \Pr(m_{i2} = 0 | x_i, \omega) \\ &\quad \times \prod_{i=r+1}^n f(y_{i1} | x_i, m_{i2} = 1, \xi) \Pr(m_{i2} = 1 | x_i, \omega). \end{aligned}$$

This representation reveals that we have no data with which to estimate directly the distribution $f(y_{i2} | x_i, y_{i1}, m_{i2} = 1, \xi)$, because all units with $m_{i2} = 1$ have y_{i2} missing. Under MAR, $f(y_{i2} | x_i, y_{i1}, m_{i2} = 1, \xi) = f(y_{i2} | x_i, y_{i1}, m_{i2} = 0, \xi)$, as discussed in Example 1.13. For MNAR models, other assumptions are needed to estimate this distribution.

The selection model formulation (15.3) for this situation is

$$f(y_i, m_{i2} | x_i, \theta, \psi) = f(y_{i1} | x_i, \theta) f(y_{i2} | x_i, y_{i1}, \theta) f(m_{i2} | x_i, y_{i1}, y_{i2}, \psi).$$

One possibility is to model $f(m_{i2} | x_i, y_{i1}, y_{i2}, \psi)$ via an additive probit or logistic regression of m_{i2} on x_i , y_{i1} , and y_{i2} . Note, however, that the coefficient of y_{i2} in this regression is not directly estimable from these data, because y_{i2} is only observed when $m_{i2} = 0$. Hence, neither pattern-mixture nor selection models are fully estimable without extra assumptions. Approaches to this problem are discussed in Sections 15.3 and 15.4.

For missingness confined to a single variable Y , John Tukey, in a discussion of Glynn et al. (1986) recorded in Holland (1986), suggested the alternative factorization

$$f(y_i, m_i | x_i, \theta, \psi) = \Pr(m_i | x_i, \psi) f(y_i | m_i = 0, \theta) \frac{\Pr(m_i | x_i, y_i, \psi)}{\Pr(m_i = 0 | x_i, y_i, \psi)},$$

which he called a “simplified selection” factorization. As Holland noted, a main advantage of this factorization is that it only involves the observed-data density, $f(y_i | x_i, m_i = 0, \theta)$, which can be estimated directly, and the missingness mechanism, $\Pr(m_i | x_i, y_i, \psi)$, which may be easy to elicit in the context of a specific application. Franks et al. (2016) develop this approach, using the term “possibly incompatible Gibbs’ Sampler” (PIGS), because it is based on distributions that may not be formally compatible, as in the chained equations approach to MI in Section 10.2.4.

Hybrids of selection and pattern-mixture models are also possible. In particular, write $m_i = (m_i^{(1)}, m_i^{(2)})$, where $m_i^{(1)}$ indexes sets of missingness patterns, and $m_i^{(2)}$ indexes individual patterns within each set. *Pattern-set mixture models* (Little 1993c) write the joint distribution of m_i and y_i in the form

$$\begin{aligned} & f\left(m_i^{(1)}, m_i^{(2)}, y_i | x_i, \xi, \psi, \omega\right) \\ &= f\left(y_i | x_i, m_i^{(1)}, \xi\right) f\left(m_i^{(2)} | x_i, y_i, m_i^{(1)}, \psi\right) f\left(m_i^{(1)} | x_i, \omega\right), \end{aligned} \quad (15.5)$$

where ξ , ψ and ω are distinct; (15.5) includes (15.3) and (15.4) as special cases where $m_i^{(1)}$ or $m_i^{(2)}$ contain just a single pattern.

Example 15.2 Pattern-Set Mixture Models for Survey Nonresponse. Unit and item nonresponse in a sample survey can be conveniently modeled using pattern-set mixture models. Write $m_i = (m_i^{(1)}, m_i^{(2)})$, where $m_i^{(1)}$ is a scalar indicator of unit nonresponse ($m_i^{(1)} = 1$ for unit nonrespondents, $m_i^{(1)} = 0$ for unit respondents), and $m_i^{(2)}$ is a vector indicator of item missingness for the survey variables. Then $m_i^{(2)} = (1, 1, \dots, 1)$ for unit nonrespondents, and $m_i^{(2)}$ has at least some components equal to zero for each unit respondent. A pattern-set mixture model for unit i is

$$\begin{aligned} & f\left(m_i^{(1)}, m_i^{(2)}, y_i | x_i, \xi, \psi, \omega\right) \\ &= f\left(y_i | x_i, m_i^{(1)}, \xi\right) f\left(m_i^{(2)} | x_i, y_i, m_i^{(1)}, \psi\right) f\left(m_i^{(1)} | x_i, \omega\right), \end{aligned} \quad (15.6)$$

where y_i are the survey variables. Equation (15.6) models unit nonresponse via a pattern-mixture model, with distributions $f(y_i|x_i, m_i^{(1)} = 1, \xi)$ for unit nonrespondents and $f(y_i|x_i, m_i^{(1)} = 0, \xi)$ for unit respondents with mixing distribution $f(m_i^{(1)}|x_i, \omega)$, and models item nonresponse for unit respondents as a selection model with components $f(y_i|x_i, m_i^{(1)} = 0, \xi)$ and $f(m_i^{(2)}|x_i, y_i, m_i^{(1)} = 0, \psi)$; the remaining distribution $f(m_i^{(2)}|x_i, y_i, m_i^{(1)} = 1, \psi)$ equals 1 when $m_i^{(2)} = (1, 1, \dots, 1)$ and zero otherwise.

A special case of this formulation that may make substantive sense is to allow unit nonresponse to be MNAR, but to assume that item nonresponse for unit respondents is MAR, when observed characteristics are sufficient to characterize differences between item respondents and item nonrespondents. In that case, the factor $f(m_i^{(2)}|x_i, y_i, m_i^{(1)} = 0, \psi) = f(m_i^{(2)}|x_i, y_{(0),i}, m_i^{(1)} = 0, \psi)$ can be ignored for likelihood inference about ψ and ω , and thus for inference about the distribution of Y .

Maximum likelihood estimates with MNAR missingness mechanisms are obtained by maximizing (15.2), and a large-sample sampling covariance matrix for the estimated parameters can be estimated using the inverse of the information matrix or the bootstrap. Explicit ML estimates can be derived in special situations, such as the pattern-mixture model in Example 15.13 below. More often, however, iterative techniques are required to maximize the likelihood, as discussed in Section 8.1 for ignorable nonresponse.

In particular, the EM algorithm has the following form for MNAR selection models: (i) find initial estimates $\theta^{(0)}, \psi^{(0)}$ of (θ, ψ) ; (ii) at iteration $t \geq 0$, given current estimates $(\theta^{(t)}, \psi^{(t)})$ of (θ, ψ) , the E step calculates

$$Q(\theta, \psi | \theta^{(t)}, \psi^{(t)}) = \int \ell(\theta, \psi | X, Y_{(0)}, Y_{(1)}, M) f(y_{(1)}|X, Y_{(0)}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}) dY_{(1)},$$

where $\ell(\theta, \psi | X, Y_{(0)}, Y_{(1)}, M)$ is the complete-data loglikelihood and $f(y_{(1)}|X, Y_{(0)}, M, \theta, \psi)$ is the density of the conditional distribution of the missing data $Y_{(1)}$, given the observed values ($X, Y_{(0)}$ and M), and parameters θ and ψ . The M step finds $\theta^{(t+1)}, \psi^{(t+1)}$ to maximize Q :

$$Q(\theta^{(t+1)}, \psi^{(t+1)} | \theta^{(t)}, \psi^{(t)}) \geq Q(\theta, \psi | \theta^{(t)}, \psi^{(t)}) \quad \text{for all } \theta, \psi.$$

Then $\theta^{(t+1)}$ and $\psi^{(t+1)}$ replace $\theta^{(t)}$ and $\psi^{(t)}$ in the next iteration of EM. By theory analogous to that in Section 8.4, each iteration increases $L(\theta, \psi | X, Y_{(0)}, M)$, and under rather general conditions, the algorithm converges to a stationary value of the full likelihood. Extensions of EM, such as ECM or PX-EM, can be helpful here, as with the ignorable case. However, for nonignorable models that

include parameters for which most of the information is missing, convergence to a maximum may be very slow. Also, particular attention needs to be paid with such models to the possibility of multiple maxima, or even ridges, of the likelihood function.

Bayesian inference for the parameters (θ, ψ) is based on their posterior distribution obtained by multiplying the full likelihood by a prior distribution for (θ, ψ) .

The remainder of this chapter is structured as follows: in Section 15.2, we give examples of models where the missingness mechanism is MNAR but *known*, in the sense that the distribution of M given X and $Y = (y_{(0)}, Y_{(1)})$ depends on fully-observed X and $Y_{(1)}$ but does not involve unknown parameters ψ . A simple example is the censored exponential sample leading to the likelihood (6.58), because values are missing when they are greater than a *known* censoring point, the distribution of M given X and Y is fully determined. Such missingness mechanisms are MNAR, but the data are *coarsened at random* (CAR), as defined in Section 6.4.

Section 15.3 considers normal models where the missingness mechanism is MNAR and depends on unknown parameters. Normal selection and pattern-mixture models are formulated, and then five approaches to the inference problem are discussed: (i) collecting data on a subsample of nonrespondents; (ii) Bayesian methods that impose a prior distribution on the parameters of the missingness mechanism; (iii) imposing restrictions on the joint model so that all the parameters have unique ML estimates; (iv) sensitivity analysis; and (v) selectively discarding data (that is, acting as if some observed values were not observed) to avoid modeling the missingness mechanism.

Section 15.4 considers some other examples of MNAR modeling, namely: models for repeated measures and categorical data, a sensitivity analysis for deviations from MAR in the chained equation models discussed in Section 10.2.4, a sensitivity analysis for survival analysis, and enhanced “tipping-point” displays.

15.2 Models with Known MNAR Missingness Mechanisms: Grouped and Rounded Data

Kulldorff (1961) discusses scoring algorithms for ML estimation from data where some units are grouped into categories. The following three examples illustrate the use of the EM algorithm in this setting. Example 15.6 describes Bayesian inference using the Gibbs’ sampler.

Example 15.3 *Grouped Exponential Sample.* Suppose the complete data are an independent random sample $(y_1, \dots, y_n)^T$ from the exponential distribution

with mean θ , but y_i is observed only for $i = 1, \dots, r < n$. The remaining $n - r$ units are grouped into J categories, such that the j th category contains values of y_i known to lie between a_j and b_j , both known, and the observed data for these $n - r$ units are counts m_j of units in the j th category, for $j = 1, \dots, J$, $\sum_{j=1}^J m_j = n - r$. This formulation includes censored data, where $a_j > 0$ and $b_j = \infty$, as well as situations where $r = 0$ and all the data are in grouped form. The coarsening mechanism is CAR, using the terminology in Section 6.4.

We expand the binary missingness indicator m_i of Section 15.1 to a variable with $J + 1$ values for this example. Specifically, let $m_i = 0$ if y_i is observed, and $m_i = j$ if y_i falls in the j th nonresponse category, that is, lies between a_j and b_j ($j = 1, \dots, J$).

The complete data belong to the regular exponential family with complete data sufficient statistic $\sum_{i=1}^n y_i$. Hence, the E step of the EM algorithm calculates, at iteration t ,

$$E\left(\sum_{i=1}^n y_i \mid Y_{(0)}, M, \theta = \theta^{(t)}\right) = \sum_{i=1}^r y_i + \sum_{j=1}^J m_j \hat{y}_j^{(t)},$$

where, from the definition of the exponential distribution,

$$\begin{aligned}\hat{y}_j^{(t)} &= E(y \mid a_j \leq y < b_j, \theta^{(t)}) \\ &= \int_{a_j}^{b_j} y \exp\left(-\frac{y}{\theta^{(t)}}\right) dy / \int_{a_j}^{b_j} \exp\left(-\frac{y}{\theta^{(t)}}\right) dy,\end{aligned}$$

and integrating by parts gives

$$\hat{y}_j^{(t)} = \theta^{(t)} + \frac{b_j e^{-b_j/\theta^{(t)}} - a_j e^{-a_j/\theta^{(t)}}}{e^{-b_j/\theta^{(t)}} - e^{-a_j/\theta^{(t)}}}. \quad (15.7)$$

The M step of EM calculates

$$\theta^{(t+1)} = n^{-1} \left(\sum_{i=1}^r y_i + \sum_{j=1}^J m_j \hat{y}_j^{(t)} \right). \quad (15.8)$$

The predicted value for a unit censored at a_j is obtained by setting $b_j = \infty$, yielding

$$\hat{y}_j^{(t)} = \theta^{(t)} + a_j.$$

If all of the $n - r$ grouped units are censored, then an explicit ML estimate can be derived. Substituting (15.7) in (15.8) yields

$$\theta^{(t+1)} = n^{-1} \left(\sum_{i=1}^r y_i + \sum_{j=1}^J m_j (\theta^{(t)} + a_j) \right).$$

Setting $\theta^{(t)} = \theta^{(t+1)} = \hat{\theta}$ and solving for θ gives

$$\hat{\theta} = r^{-1} \left(\sum_{i=1}^r y_i + \sum_{j=1}^J m_j a_j \right).$$

In particular, if $a_j = c$ for all j , that is, the units have a common censoring point, then

$$\hat{\theta} = m^{-1} \left(\sum_{i=1}^m y_i + (n-m)c \right),$$

which is the estimate derived directly in Example 6.29.

Example 15.4 *Grouped Normal Data with Covariates.* Suppose that data on a normally distributed outcome variable Y are grouped as in Example 15.3, where unit i is classified in group j if it is known to lie between a_j and b_j , but now the complete Y values are independent with linear regressions on fully observed covariates X_1, X_2, \dots, X_p . That is, y_i is normal with mean $\beta_0 + \sum_{k=1}^p \beta_k x_{ik}$ and constant variance σ^2 . The complete-data sufficient statistics are $\sum y_i$, $\sum y_i x_{ik}$ ($k = 1, \dots, p$) and $\sum y_i^2$. Hence, the E step of the EM algorithm computes

$$E \left(\sum_{i=1}^n y_i \mid Y_{(0)}, M, \theta = \theta^{(t)} \right) = \sum_{i=1}^r y_i + \sum_{i=r+1}^n \hat{y}_i^{(t)},$$

$$E \left(\sum_{i=1}^n y_i x_{ik} \mid Y_{(0)}, M, \theta = \theta^{(t)} \right) = \sum_{i=1}^r y_i x_{ik} + \sum_{i=r+1}^n \hat{y}_i^{(t)} x_{ik}, \quad k = 1, 2, \dots, p,$$

$$E \left(\sum_{i=1}^n y_i^2 \mid Y_{(0)}, M, \theta = \theta^{(t)} \right) = \sum_{i=1}^r y_i^2 + \sum_{i=r+1}^n \hat{y}_i^{(t)2} + \hat{s}_i^{(t)2},$$

where $\theta = (\beta_0, \beta_1, \dots, \beta_p, \sigma^2)$, $\theta^{(t)} = (\beta_0^{(t)}, \beta_1^{(t)}, \dots, \beta_p^{(t)}, \sigma^{(t)2})$ is the current estimate of θ , $\hat{y}_i^{(t)} = \mu_i^{(t)} + \sigma^{(t)} \delta_i^{(t)}$, $\hat{s}_i^{(t)2} = \sigma^{(t)2} (1 - \gamma_i^{(t)})$, $\mu_i^{(t)} = \beta_0^{(t)} + \sum_{k=1}^p \beta_k^{(t)} x_{ik}$,

and $\delta_i^{(t)}$ and $\gamma_i^{(t)}$ are corrections for the MNAR nonresponse; these corrections take the form

$$\delta_i^{(t)} = -\frac{\phi(d_i^{(t)}) - \phi(c_i^{(t)})}{\Phi(d_i^{(t)}) - \Phi(c_i^{(t)})},$$

$$\gamma_i^{(t)} = \delta_i^{(t)2} + \frac{d_i^{(t)}\phi(d_i^{(t)}) - c_i^{(t)}\phi(c_i^{(t)})}{\Phi(d_i^{(t)}) - \Phi(c_i^{(t)})},$$

where ϕ and Φ are the standard normal density and cumulative distribution functions respectively, and for units i in the j th category ($M_i = j$ or, equivalently, $a_j < Y \leq b_j$),

$$c_i^{(t)} = (a_j - \mu_i^{(t)}) / \sigma^{(t)} \quad \text{and} \quad d_i^{(t)} = (b_j - \mu_i^{(t)}) / \sigma^{(t)}.$$

The M step calculates the regression of Y on X_1, \dots, X_p using the expected values of the complete-data sufficient statistics found in the E step. This model is applied to a regression of log(blood lead) using grouped data in Hasselblad et al. (1980).

Example 15.5 Censored Normal Data with Covariates (Tobit Model). The Tobit model in the econometric literature (Amemiya 1984), named after an earlier econometric application (Tobin 1958), is a special case of the previous example where positive values of Y are fully recorded, but negative values are censored (that is, lie in the interval $(-\infty, 0)$). In the notation of Example 15.4, all observed y_i are positive, $J=1$, $a_1 = -\infty$, and $b_1 = 0$. For the E step of EM, for censored units $c_i^{(t)} = -\infty$, $d_i^{(t)} = -\mu_i^{(t)} / \sigma^{(t)}$, $\delta_i^{(t)} = -\phi(d_i^{(t)}) / \Phi(d_i^{(t)})$, $\gamma_i^{(t)} = \delta_i^{(t)}(\delta_i^{(t)} + \mu_i^{(t)} / \sigma_i^{(t)})$. Hence,

$$\hat{y}_i^{(t)} = E(y_i | \theta^{(t)}, x_i, y_i \leq 0) = \mu_i^{(t)} - \sigma^{(t)} \lambda(-\mu_i^{(t)} / \sigma^{(t)}),$$

where $\lambda(z) = \phi(z) / \Phi(z)$ (the inverse of the so-called Mills ratio), and $-\sigma^{(t)} \lambda(-\mu_i^{(t)} / \sigma^{(t)})$ is the correction for censoring. Substituting ML estimates of the parameters yields the predicted values

$$\hat{y}_i^{(t)} = E(y_i | \hat{\theta}, x_i, y_i \leq 0) = \hat{\mu}_i - \hat{\sigma} \lambda(-\hat{\mu}_i / \hat{\sigma}),$$

for censored units, where $\hat{\mu}_i = \hat{\beta}_0 + \sum_{k=1}^p \hat{\beta}_k x_{ik}$.

Example 15.6 Multiple Imputation of Coarsened Data from the Health and Retirement Survey (HRS). Survey questions concerning household financial variables can be subject to high rates of missing data. One partial solution is to

use questions that bracket amounts within intervals (e.g., \$5000–9999) whenever the respondent refuses or is unable to provide an exact response to a question. These “bracketed response” formats significantly reduce the rates of completely missing data for financial variables, but yield coarsened data that are a mixture of actual reported responses, bracketed (or interval-censored) replies, and completely missing data. Heitjan and Rubin (1990) applied MI to a related problem involving age heaping. Heeringa et al. (2002) develop multiple imputations of coarsened and missing data for 12 asset and liability variables in the health and retirement survey (HRS), based on an extension of the general location model of Section 14.2 suggested in Little and Su (1987).

We present the model for bivariate data for simplicity, but it extends directly to more than two variables. Let $y_i = (y_{i1}, y_{i2})$ denote two nonnegative asset or liability measures for unit i , and let $t_i = (t_{i1}, t_{i2})$ indicate the existence of positive amounts: $t_{ij} = 1$ if $y_{ij} > 0$, $t_{ij} = 0$ if $y_{ij} = 0$. Also, to allow positive holdings to be log-normal, let $z_i = (z_{i1}, z_{i2})$ be partly observed variables such that

$$y_i = \begin{cases} (\exp(z_{i1}), \exp(z_{i2})), & \text{if } t_i = (1, 1), \\ (\exp(z_{i1}), 0), & \text{if } t_i = (1, 0), \\ (0, \exp(z_{i2})), & \text{if } t_i = (0, 1), \\ (0, 0), & \text{if } t_i = (0, 0), \end{cases} \quad (15.9)$$

and

$$(z_i | t_i = (j, k), \theta) \sim N_2(\mu_{jk}, \Sigma). \quad (15.10)$$

The exponential transformations in (15.9) imply that the nonzero assets and liabilities are log-normal, reflecting right skewness of their distributions. The model (15.10) allows distinct means $\{\mu_{jk}\}$ for each pattern (j, k) of zero/nonzero amounts, but assumes a constant covariance matrix Σ , as in the general location model of Section 14.2. Note that it is unrealistic to apply (15.10) directly to y_i , because the positive values are skewed, and the model assumption of a constant covariance matrix is untenable – for example, the variance of a component y_{ij} is zero when $t_{ij} = 0$. The means of the unobserved components of z_i in cells j with $t_{ij} = 0$ do not affect y_i and are constrained to zero; however, an alternative approach that may speed the convergence to ML estimates is to treat them as parameters to be estimated by the algorithm, as in the PX-EM algorithm of Section 8.5.3.

Heeringa et al. (2002) apply this model assuming that t_i is fully observed, that is, that the household’s ownership (yes/no) of each net-worth component is always known. Methods can also be developed for the situation where some components of t_i are missing. Individual components y_{ij} can be observed, completely missing, or known to lie in an interval, say (l_{ij}, u_{ij}) . An attractive feature of Gibbs’ sampling is that draws of the missing values can be generated one

variable at a time, conditioning on current draws of the parameters, as well as the observed or drawn values of all the other variables. Because the conditional distribution of any one variable given the others is normal, interval-censored information about that variable is easily incorporated in the draws; the equations parallel the E step of Example 15.4, but with draws replacing conditional means.

Multiple-imputation large-sample inference methods described in Section 10.2 yield point estimates, interval estimates, and test statistics for model parameters. Imputations in the highest open-ended dollar categories are highly sensitive to model specification, and in Heeringa et al. (2002), the imputations are truncated so that they do not exceed the largest recorded value in the dataset. Extensions of the log-normal model described here might be developed to fit the tails of the distributions more closely.

Table 15.1 shows the results of applying this method to the problem of estimating the mean and quantiles of total net worth for the HRS survey population, found by aggregating 23 net worth component variables. The following methods are included for comparison purposes:

Complete-case analysis: The distribution of total net worth was estimated from the 4566 (60.0%) of the 7607 cooperating HRS Wave 1 households who provided complete information on holdings and amounts for each of the 23 asset and liability components.

Mean imputation: The mean of those observed values that fell within a bracket was imputed for all units with missing values in that bracket. If a value of a

Table 15.1 Example 15.6, health and retirement survey wave 1 estimates of the distribution of total household net worth, in thousands of dollars

Estimand	MI bayes		Complete units		Mean imputation		MI hot deck	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Mean	247.9	10.6	186.8	9.1	213.5	7.7	232.5	9.4
SD	598.6	77.2	417.3	27.5	443.7	29.4	491.1	42.8
Q25	28.9	2.2	15.3	2.8	28.4	2.0	29.5	2.4
Q50	99.7	4.4	78.0	3.2	97.3	4.4	100.8	5.0
Q75	240.1	9.8	195.5	10.0	218.0	7.0	240.4	8.6
Q90	537.1	25.2	408.5	15.3	471.6	23.6	515.0	30.9
Q95	902.5	54.7	663.0	28.0	779.6	33.1	839.8	56.6
Q99	2 642.3	264.1	1 995.0	107.1	2 142.1	62.035	2 317.8	216.0
Max	15 663	9 458	6 202	322.0	9 096	469.4	9 645	3 070

variable was completely missing (no bracket information), the overall mean of the observed values for that variable was imputed for the unit.

Multiple imputation based on a univariate hot deck: A univariate hot deck method was originally used to produce a single imputation for item missing data in the HRS Wave 1 data set, where each asset and liability variable was imputed independently. All observed and missing units were assigned to hot deck cells based on covariate information, which included age, race, sex, marital status of the household head, and bracket boundaries when bracketed information was available. Each missing value was then imputed using the observed value of a randomly selected observed unit within the same hot deck cell. Repeating this hot-deck procedure with different random donors chosen within each adjustment cell created 20 multiply-imputed data sets.

The summaries of the estimated distributions in Table 15.1 incorporate sampling weights, and for the hot deck and Bayes methods are averaged over the 20 multiply-imputed data sets. Complete-data standard errors were estimated using the Jackknife Repeated Replications (JRRs) method (Wolter 1985), and they reflect the influences of weighting, stratification, and clustering of the complex multistage HRS sample design. Standard errors for the mean-imputation methods do not account for imputation uncertainty. Standard errors for the univariate hot deck, Bayes' and sequential regression methods were computed using the multiple-imputation formulae in Section 10.2, with the within-imputation variance being the design-based JRR variance estimate. These variances incorporate estimates of imputation uncertainty, as well as the effects of the complex sample design.

It can be seen from Table 15.1 that (i) complete-case (CC) analysis appears to underestimate markedly the distribution of household net worth for HRS households; (ii) compared to stochastic imputation alternatives, mean substitution also appears to underestimate the mean and percentiles of the full net-worth distribution. The standard deviation of the imputed household net-worth distribution produced by this deterministic imputation method is attenuated when compared to the standard deviations in net-worth amounts implied by the stochastic hot deck and Bayesian alternatives; (iii) the hot-deck method produces lower estimated values for the mean and upper quantiles of the distribution than the Bayesian method. This finding may relate to the fact that, unlike the Bayes method, the hot-deck imputations do not utilize information (including bracketing) for other variables in the multivariate vector of net-worth components. Analyses of the fraction of missing information indicate that statistics most influenced by the upper tails of the component distributions, namely the mean, standard deviation, Q99 and maximum value, have the highest degree of imputation uncertainty.

15.3 Normal Models for MNAR Missing Data

15.3.1 Normal Selection and Pattern-Mixture Models for Univariate Missingness

In this section, we assume the complete data are a random sample (y_i, x_i) , $i = 1, \dots, n$ on a continuous variable Y , and a set of covariates, X , which are fully observed. We suppose that $\{y_i, i = 1, \dots, r\}$ are observed and $\{y_i, i = r + 1, \dots, n\}$ are missing; m_i is the missingness indicator for y_i , taking values $m_i = 0$ for $i = 1, \dots, r$ and $m_i = 1$ for $i = r + 1, \dots, n$. We describe selection and pattern-mixture models for this data structure.

Example 15.7 A Probit Selection Model for Univariate Missingness. Heckman (1976) proposed the following selection model (15.3): for unit i ,

$$\begin{aligned} (y_i | x_i, \theta, \psi) &\sim_{\text{ind}} N(\beta_0 + \beta_1 x_i, \sigma^2), \\ (m_i | x_i, y_i, \theta, \psi) &\sim_{\text{ind}} \text{BERN}(\Phi(\psi_0 + \psi_1 x_i + \psi_2 y_i)), \end{aligned} \quad (15.11)$$

where $\theta = (\beta_0, \beta_1, \sigma^2)$, BERN denotes the Bernoulli distribution, and Φ denotes the probit (cumulative normal) distribution function. Y is the outcome variable in a regression model, but (15.11) could also be used to model a predictor variable with missing values. Greenlees et al. (1982) considered an analogous model to (15.11), but with a logistic rather than a probit model for the conditional distribution of m_i given x_i and y_i .

Note that if $\psi_2 = 0$, then the missing data are MAR. If $\psi_2 \neq 0$, then the missing data are MNAR, because missingness of Y depends on the value of Y , which is missing for nonrespondents. Heckman (1976) used a two-step least squares method to fit this model. Alternatively, ML estimates can be obtained, for example by applying the EM algorithm, treating the unobserved values of Y as missing data. Details of the algorithm are omitted.

The main obstacle in fitting the model is lack of information about ψ_2 , the coefficient of y_i in the distribution for m_i in (15.11). The assumption of normality of residuals in the unrestricted population in principle provides information, in that lack of normality in the distribution of observed residuals in the complete units is then evidence of a MNAR mechanism; however, relying on normality to estimate ψ_2 is highly questionable, because we rarely know with any certainty that the distribution of y_i given x_i is normal (Little 1985a; Little and Rubin 2002, chapter 15). Better approaches to the lack of information about ψ_2 are discussed in Section 15.3.

Example 15.8 A Normal Pattern-Mixture Model for Univariate Missingness. An alternative to (15.11) is the following pattern-mixture model (15.4):

$$\begin{aligned} (y_i|m_i = m, x_i, \xi, \omega) &\sim_{\text{ind}} N(\beta_0^{(m)} + \beta^{(m)}x_i, \sigma^{(m)2}), \\ (m_i|x_i, \xi, \omega) &\sim_{\text{ind}} \text{BERN}(\Phi(\omega_0 + \omega_1 x_i)), \end{aligned} \quad (15.12)$$

where $\xi = (\beta_0^{(m)}, \beta^{(m)}, \sigma^{(m)2}, m = 0, 1)$. This model implies that the distribution of y_i given x_i in the whole population is a mixture of two normal distributions, with mean

$$[1 - \Phi(\omega_0 + \omega_1 x_i)] [\beta_0^{(0)} + \beta^{(0)} x_i] + [\Phi(\omega_0 + \omega_1 x_i)] [\beta_0^{(1)} + \beta^{(1)} x_i].$$

The parameters $(\beta_0^{(0)}, \beta^{(0)}, \sigma^{(0)2}, \phi)$ in this model can be estimated from the data, but the parameters $(\beta_0^{(1)}, \beta^{(1)}, \sigma^{(1)2})$ are not estimable, because y_i is missing when $m_i = 1$. When the data are MAR, the distribution of Y given X is the same for units with Y observed and for units with Y missing (see Example 1.13) and $\beta_0^{(1)} = \beta_0^{(0)} = \beta_0$, $\beta^{(1)} = \beta^{(0)} = \beta$, $\sigma^{(1)2} = \sigma^{(0)2} = \sigma^2$. When the data are not MAR, other assumptions are needed to allow estimates of $(\beta_0^{(1)}, \beta^{(1)}, \sigma^{(1)2})$.

The number of inestimable parameters in this model can be reduced by assuming that the distributions of Y for units observed and missing Y differ only in their intercepts, that is

$$\begin{aligned} (y_i|m_i = m, x_i, \xi, \omega) &\sim_{\text{ind}} N(\beta_0^{(m)} + \beta x_i, \sigma^{(m)2}), \\ (m_i|x_i, \xi, \omega) &\sim_{\text{ind}} \text{BERN}(\Phi(\omega_0 + \omega_1 x_i)). \end{aligned} \quad (15.13)$$

The difference in means for nonrespondents and respondents, namely the effect $\delta = \beta_0^{(1)} - \beta_0^{(0)}$, characterizes the differences between Y for respondents and nonrespondents in this model.

We make some general comments about the selection and pattern-mixture formulations of the model in Examples 15.7 and 15.8:

1. Both approaches to modeling the joint distribution of Y and M are legitimate, in the sense that neither can be ruled out empirically.
2. The selection model formulation (Example 15.7) was used to characterize MAR in Chapter 6. It is more natural when substantive interest concerns the relationship between Y and X in the whole population, and strata defined by the pattern of missing data are not substantively meaningful. However, parameters of the whole population can be derived from the pattern-mixture model by averaging parameters over the missingness patterns.

3. The pattern-mixture model (Example 15.8) is arguably easier to explain to subject-matter experts than the selection model. In particular, in (15.13), the difference between respondents and nonrespondents is characterized by δ , a parameter that has a simple interpretation as a difference in means. In (15.11), the difference between respondents and nonrespondents is characterized by the parameter ψ_2 , which has the more obscure interpretation as the effect of increasing Y by one unit on the probit of the probability of non-response, holding the covariates X constant. In the Greenlees et al. (1982) model, the probit is replaced by the log-odds, but the interpretation of ψ_2 remains difficult.
4. The pattern-mixture model is often easy to fit, given assumptions to render the parameters estimable. Also, imputations of the missing values are based on the predictive distribution of Y given X and $M=0$, which is modeled directly in the pattern-mixture factorization.

Considerations 3 and 4 lead us to favor pattern-mixture models, particularly when we consider sensitivity analyses as in Section 15.3.6. The pattern-mixture approach is also favored in Carpenter and Kenward (2014).

In the following subsections, five approaches that address the lack of estimability of certain parameters in these models are distinguished:

- (a) Follow up a sample of nonrespondents, and incorporate this information into the main analysis.
- (b) Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the data provide information for all the parameters, although inferences tend to be sensitive to the choice of prior distribution.
- (c) Impose additional restrictions on model parameters, such as on the regression coefficients in (15.11)–(15.13).
- (d) Conduct analysis to assess sensitivity of inferences for quantities of interest to different choices of the values of parameters poorly estimated from the data.
- (e) Selectively discard data to avoid modeling the missingness mechanism.

We now examine each of these approaches in turn.

15.3.2 Following up a Subsample of Nonrespondents

One way to reduce sensitivity of inference to MNAR nonresponse is to follow up at least some nonrespondents to obtain the desired information. Even if only a few nonrespondents are followed up, these can be exceedingly helpful in reducing sensitivity of inference, as the following simulation experiment illustrates.

Example 15.9 Decreased Sensitivity of Inference with Follow-ups. Glynn et al. (1986) performed a series of simulations using normal and lognormal data, which were used to study the decreased sensitivity of inference when follow-up data are obtained from nonrespondents. For the normal data, a sample of 400 standard normal deviates was drawn from an essentially infinite population, and the logistic nonresponse mechanism

$$\Pr(m_i = 1|y_i) = [1 + \exp(1 + y_i)]^{-1}$$

was applied to create 101 nonrespondents. Then, various fractions of the 101 nonrespondents were randomly sampled to create follow-up data among initial nonrespondents. The resultant data consisted of (y_i, m_i) for respondents and followed-up nonrespondents, but only m_i was observed for nonrespondents who were not followed up.

Two models were used to analyze the data. First, the pattern-mixture model (15.12) without covariates was used with the prior distribution on $(\mu_{(0)}, \mu_{(1)}, \ln \sigma_{(0)}, \ln \sigma_{(1)}, \pi)$ proportional to a constant. Second, the data were analyzed under the correct normal/logistic response selection model:

$$(y_i|\mu, \sigma^2) \sim N(\mu, \sigma^2), \\ \text{logit}(\Pr(m_i = 1|y_i, \alpha_0, \alpha_1)) = \alpha_0 + \alpha_1 y_i,$$

where the improper prior distribution on $(\mu, \alpha_0, \alpha_1, \ln \sigma)$ was proportional to a constant. The entire simulation was repeated with a different data set, with 400 lognormal values (exponentiated standard normal deviates) and 88 nonrespondents created using the MNAR logistic missingness mechanism $\Pr(m_i = 1|y_i) = 1/(1 + \exp(1 + y_i))$. Again, various fractions of the nonrespondents were randomly sampled to create follow-up data among the nonrespondents. The same two models used to analyze the normal data were applied to the lognormal data. Note that, whereas for the normal data the selection model was correct and the pattern-mixture model incorrect, for the lognormal data both models are incorrect.

Table 15.2 summarizes the generated data, both normal and lognormal. Table 15.3 gives estimates of the population means for both models with both the normal and lognormal data. Several trends are readily apparent: first, the mixture model appears to be somewhat more robust than the selection model, doing as well as the selection model when the selection model is correct and doing better than the selection model when neither is correct. Second, the larger the fraction of follow-ups, the better the estimates under both models; with full follow-up, the estimates from the two models are very similar, differing only in their precision (not displayed here). Third, using the mixture model with even a few follow-ups yields reasonable estimates. Glynn et al. (1986) use

Table 15.2 Example 15.9, sample moments of generated data^a

	Normal $N(0,1)$ data			Lognormal $\exp[N(0,1)]$ data		
	N	Mean	SD	N	Mean	SD
Respondents	299	0.150	0.982	312	1.857	2.236
Nonrespondents	101	-0.591	0.835	88	0.724	0.571
Total	400	-0.037	1.000	400	1.608	2.047
Population values		0.0	1.0		1.649	2.161

^aNormal data are sampled from the normal (0,1) distribution; lognormal data are the exponentiated normal values. Response is determined by a logistic response function: $\Pr(m_i = 1 | y_i) = [\alpha_0 + \exp(1 + \alpha_1 y_i)]^{-1}$, where $(\alpha_0, \alpha_1) = (1, 1)$ for normal data and (0,1) for lognormal data.

MI to draw inferences from survey data of retired men with follow-ups, using an extension of the mixture model that includes covariates.

15.3.3 The Bayesian Approach

Example 15.10 *Inference about the Sample Mean with MNAR Nonresponse, in the Presence of Covariates.* Rubin (1977) applies a Bayesian approach to the following slight modification of the pattern-mixture model in Example 15.8:

$$(y_i | m_i = m, x_i, \xi, \omega) \sim_{\text{ind}} N(\beta_0^{(m)} + \beta_1^{(m)}(x_i - \bar{x}_0), \sigma^2), \quad (15.14)$$

Table 15.3 Example 15.9, estimates of population mean using respondent data of Table 15.2 and follow-up data from some randomly selected nonrespondents

Normal $N(0,1)$ data (population mean = 0)			Lognormal $\exp[N(0,1)]$ data (population mean = 1.649)		
Number of follow-ups (of 101)	Mixture model	Selection model	Number of follow-ups (of 88)	Mixture model	Selection model
11	-0.010	-0.009	9	1.58	0.934
24	-0.025	-0.029	21	1.60	1.030
28	-0.006	-0.008	25	1.61	1.054
101	-0.037	-0.037	88	1.61	1.605

with the following prior distributions on the parameters:

$$\begin{aligned} p(\beta_0^{(0)}, \beta_1^{(0)}, \log \sigma^2) &\propto \text{const.}, \\ p(\beta_1^{(1)} | \beta_0^{(0)}, \beta_1^{(0)}, \sigma^2) &\sim N_q(\beta_1^{(0)}, \psi_1^2 \beta_1^{(0)} \beta_1^{(0)\top}), \\ p(\beta_0^{(1)} | \beta_1^{(1)}, \beta_0^{(0)}, \beta_1^{(0)}, \sigma^2) &\sim N(\beta_0^{(0)}, \psi_2^2 (\beta_0^{(0)})^2), \end{aligned}$$

where $\beta_0^{(0)}$ and $\beta_1^{(1)}$ are parameters representing the adjusted means of Y in the respondent and nonrespondent populations at the respondent covariate mean $\bar{x}_{(0)}$. The parameter ψ_1 measures *a priori* uncertainty about the regression slope coefficients, and the parameter ψ_2 measures uncertainty in the adjusted mean. The missingness mechanism is ignorable for likelihood-based inferences if $\psi_1 = \psi_2 = 0$.

Let $\bar{y}_{(0)}$ and $\bar{x}_{(0)}$ denote the respondent means of Y and X . The posterior distribution of \bar{y} is normal with mean $\bar{y}_{(0)} + \hat{\beta}_1^{(0)}(\bar{x} - \bar{x}_{(0)})$, namely the regression estimator, and variance $\bar{y}_{(0)}^2(\psi_1^2 h_1^2 + \psi_2^2 h_2^2 + h_3^2)$, where

$$\begin{aligned} h_1^2 &= \left(\sigma^2 / \bar{y}_{(0)}^2 \right) \left[\left(\hat{\beta}_1^{(0)}(\bar{x} - \bar{x}_{(0)}) \right)^2 / \sigma^2 + (\bar{x} - \bar{x}_{(0)})^T S_{xx}^{-1} (\bar{x} - \bar{x}_{(0)}) \right] \\ h_2^2 &= p^2 \{1 + \sigma^2 / (r\bar{y}_{(0)})\} \end{aligned}$$

where r is the number of respondents and p is the proportion of missing values, and

$$h_3^2 = \left(\sigma^2 / \bar{y}_{(0)}^2 \right) ((p/r) + (\bar{x} - \bar{x}_{(0)})^T S_{xx}^{-1} (\bar{x} - \bar{x}_{(0)})).$$

Here S_{xx} is the sum-of-squares and cross-products matrix of X for respondents. The width of the associated large sample 95% posterior probability interval, $3.92\bar{y}_{(0)}(\psi_1^2 h_1^2 + \psi_2^2 h_2^2 + h_3^2)^{1/2}$, involves three components. The first, $\psi_1^2 h_1^2$, is the relative variance due to uncertainty about the equality of the slopes of Y on X in the respondent and nonrespondent groups. The term $\psi_2^2 h_2^2$ reflects uncertainty about the equality of the Y means for respondents and nonrespondents at $X = \bar{x}_{(0)}$. The term h_3^2 represents the uncertainty introduced by nonresponse that is present even when the respondent and nonrespondent distributions are equal, that is, when $\psi_1 = \psi_2 = 0$, so that the missingness mechanism is ignorable.

Rubin (1977) illustrates the method of Example 15.10 with data from a survey of 660 schools, 472 of which filled out a compensatory reading questionnaire consisting of 80 items. Twenty-one dependent variables (Y 's) and 35 background variables (X 's) describing the school and the socioeconomic status and achievement of the students were considered. The dependent variables in the study measure characteristics of compensatory reading in the form of

Table 15.4 Example 15.10, widths of subjective 95% intervals of finite population means \bar{y} , as percentages of observed sample mean $\bar{y}_{(0)}$

Variable	$\psi_1 = 0$				$\psi_1 = 0.4$			
	$\psi_2 = 0$	$\psi_2 = 0.1$	$\psi_2 = 0.2$	$\psi_2 = 0.4$	$\psi_2 = 0$	$\psi_2 = 0.1$	$\psi_2 = 0.2$	$\psi_2 = 0.4$
17B	5.6	8.0	12.7	23.7	6.0	8.3	12.9	23.6
18A	7.9	9.8	13.9	24.2	8.1	9.9	14.0	24.3
18B	15.4	16.5	19.3	27.8	16.6	17.6	20.2	28.5
23A	2.1	6.1	11.6	22.9	2.3	6.1	11.6	22.9
23C	2.0	6.0	11.6	22.9	2.0	6.1	11.6	22.9
32A	1.2	5.8	11.5	22.8	1.2	5.8	11.5	22.8
32D	1.1	5.8	11.4	22.8	1.1	5.8	11.4	22.8

Description of outcome variables:

17B: Compensatory reading carried out during school hours released from other classwork.

18A: Compensatory reading carried out during time released from social studies, science, and/or foreign language.

18B: Compensatory reading carried out during time released from mathematics.

23A: Frequency of organizing compensatory reading class into groups by reading grade level.

23C: Frequency of organizing compensatory reading class into groups by shared interests.

32A: Compensatory reading teaches textbooks other than basal readers.

32D: Compensatory reading teaches teacher-prepared materials.

frequency with which they were present, and were scaled to lie between zero (never) and one (always).

Table 15.4 shows the width of the large sample 95% interval for \bar{y} for seven of these outcome variables, expressed as a percentage of the observed mean, as a function of ψ_1 and ψ_2 . The uncertainty about equality of the slopes of the regressions for respondents and nonrespondents, modeled by the quantity ψ_1 , has a negligible impact on the interval, reflecting low values of h_1 . The values of h_2 are only marginally greater than the proportion of missing values, $p = 0.2848$. Thus, the contribution to the interval width of uncertainty about equality of the adjusted means in the respondent and nonrespondent populations is represented by $4h_2\psi_2 \approx 4p\psi_2 = 1.14\psi_2$.

The quantity ψ_2 has a major impact on the interval widths. For example, the effect of increasing the value of ψ_2 from 0 to 0.1 is to triple the interval widths in variables 23A and 23C, and to increase the interval widths in variables 32A and 32D by a factor of 5. On the other hand, for variables 17B, 18A, and in particular 18B, the component attributable to residual variance from the regression, h_3 , is more pronounced, although the other component is still nonnegligible for $\psi_2 \geq 0.1$. The example illustrates dramatically the potential impact of nonresponse bias, and the extent to which it is dependent on quantities (such as ψ_2) that generally cannot be estimated from the data at hand.

15.3.4 Imposing Restrictions on Model Parameters

Parameters for MNAR models, where some parameters are not necessarily estimable, can be made estimable by imposing model restrictions that reduce the number of parameters. The success of this approach rests heavily on whether these assumptions are realistic and well justified. We provide three examples: in Example 15.11, the assumptions are questionable, and the resulting estimates appear to be biased, to judge from comparisons with external data; whereas in Examples 15.12 and 15.14, the assumptions may have more reasonable bases.

Example 15.11 *Income Nonresponse in the US Current Population Survey.*

Lillard et al. (1982, 1986) apply the model (15.11) of Example 15.7 to income nonresponse in four rounds of the Current Population Survey (CPS) Income Supplement, conducted in 1970, 1975, 1976, and 1980. In 1980, their sample consisted of 32 879 employed white civilian males aged 16–65 who reported receipt (but not necessarily amount) of W =wages and salary earnings and who were not self-employed. Of these individuals, 27 909 reported the value of W and 4970 did not. In the notation of Example 15.7, Y_1 is defined to equal $(W^{\gamma-1})/\gamma$, where γ represents a power transformation of the kind proposed in Box and Cox (1964). The predictors X were chosen as Education (six categories), Years of market experience (four linear splines, Exp 0–5, Exp 5–10, Exp 10–20, Exp 20+), Probability of being in first year of market experience (Prob 1), Region (South or other), Child of household head (1 = yes, 0 = no), Other relative of household head or member of secondary family (1 = yes, 0 = no), Personal interview (1 = yes, 0 = no), and Year in survey (1 or 2).

The last four variables were omitted from the earnings equation; that is their coefficients in the vector β were set equal to zero. The variables Education, Years of market experience, and Region were omitted from the response equation; that is their coefficients in the vector ψ were set to zero.

Most empirical studies model the logarithm of earnings, the transformation obtained by letting $\gamma \rightarrow 0$. Table 15.5 shows estimates of the regression coefficients β for this transformation of earnings, calculated, first by ordinary least squares (OLS) on the respondents, a procedure that effectively assumes ignorable nonresponse ($\psi_2 = 0$); and second, by ML for the model (15.11) of Example 15.7. In this table, ρ is the correlation between Y and the normal latent variable in the Heckman (1976) model, and is related to the parameter ψ_2 in (15.11) by the expression $\rho = \psi_2\sigma/\sqrt{1 + \psi_2^2\sigma^2}$. For the ML procedure applied to these data, $\hat{\rho} = -0.6812$ implying higher income amounts for nonrespondents than are predicted under MAR. The regression coefficients from OLS and ML in Table 15.5 are quite similar, but the difference in intercepts ($9.68 - 9.50 = 0.18$ on the log scale) between ML and OLS implies a roughly 20% increase in

Table 15.5 Example 15.11, estimates for the regression of $\ln(\text{earnings})$ on covariates: 1980 Current Population Survey

Variable	OLS on 27 909 respondents	Coefficient (β_1) from ML for selection model on 32 879 survey units
Constant	9.5013 (0.0039)	9.6816 (0.0051)
Sch 8	0.2954 (0.0245)	0.2661 (0.0202)
Sch 9–11	0.3870 (0.0206)	0.3692 (0.0169)
Sch 12	0.6881 (0.0188)	0.6516 (0.0158)
Sch 13–15	0.7986 (0.0201)	0.7694 (0.0176)
Sch 16+	1.0519 (0.0199)	1.0445 (0.0178)
Exp 0–5	−0.0225 (0.0119)	−0.0294 (0.0111)
Exp 5–10	0.0534 (0.0038)	0.0557 (0.0039)
Exp 10–20	0.0024 (0.0016)	0.0240 (0.0016)
Exp 20+	−0.0052 (0.0008)	−0.0036 (0.0008)
Prob 1	−1.8136 (0.1075)	−1.7301 (0.0945)
South	−0.0654 (0.0087)	−0.0649 (0.0085)
$\rho = \psi_2 \sigma / \sqrt{1 + \psi_2^2 \sigma^2}$	0	−0.6842

the predicted income amounts for MNAR nonresponse, a very substantial correction.

Lillard et al. (1982) fit their stochastic censoring model for a variety of other choices of γ . Table 15.6 shows the maximized loglikelihood for three values of γ , namely, 0 (the log model), 1 (the model for raw income amounts), and 0.45, the ML estimate of γ for a random subsample of the data. The maximized loglikelihood at $\hat{\gamma} = 0.45$ is much larger than that at $\gamma = 0$ or at $\gamma = 1$, indicating that the normal selection models for raw and log-transformed earnings are not

Table 15.6 Example 15.11, the maximized loglikelihood as a function of γ with associated values of $\hat{\rho}$

γ	Maximized loglikelihood	$\hat{\rho}$
0	−300 613.4	−0.681 2
0.45	−298 169.7	−0.652 4
1.0	−300 563.1	0.856 9

Source: Lillard et al. (1982). Reproduced with permission of Chicago Press.

supported by the data. Table 15.6 also shows values of $\hat{\rho}$ as a function of γ . Note that for $\gamma = 0$ and $\gamma = 0.45$, $\hat{\rho}$ is negative, the distribution of respondent income residuals is left-skewed, and large income amounts for nonrespondents are needed to fill the right tail. On the other hand, when $\gamma = 1$, $\hat{\rho}$ is positive, the distribution of respondent residuals is right-skewed, and small income amounts for nonrespondents are needed to fill the left tail. Thus, the table reflects sensitivity of the correction to skewness in the transformed-income respondent residuals.

Lillard et al.'s (1982) best-fitting model, with $\hat{\gamma} = 0.45$, predicts income amounts for nonrespondents that are 73% larger on average than imputations supplied by the Census Bureau, which uses a hot deck method that assumes ignorable nonresponse. However, as Rubin (1983b) notes, this large adjustment is founded on the normal assumption for the population residuals from the $\gamma = 0.45$ model. It is quite plausible that nonresponse is MAR, and the unrestricted residuals follow the same (skewed) distribution as that in the respondent sample. Indeed, comparisons of Census Bureau imputations with Internal Revenue Service (IRS) income amounts from matched CPS/IRS files do not indicate substantial underestimation from these hot-deck imputations (David et al. 1986).

Example 15.12 *Estimating Incidence of Acquired Immune Deficiency Syndrome (AIDS) from a Demographic Survey with Randomly Assigned Interviewers.* Population-based surveys were considered the gold standard for estimating human immunodeficiency virus (HIV) prevalence in the 1990s. However, prevalence rates based on representative surveys may be biased because of non-response, with refusal to participate in the HIV test plausibly related to HIV status, even after controlling for other observed variables. Heckman-type selection models have been applied to adjust HIV prevalence for this MNAR mechanism, but the restrictions on the regression coefficients needed to estimate these models are often questionable. In contrast, Janssens et al. (2014) exploited random assignment of interviewers to households in the sample to estimate the participation selection effect, for a survey of 1992 individuals in urban Namibia that included an HIV test. Specifically, their model took the form:

$$\begin{aligned}\Pr(y_i = 1|x_i, z_i, \beta, \psi) &= \Phi(\beta_0 + \beta_1 x_i), \\ \Pr(m_i = 1|y_i, x_i, z_i, \beta, \psi) &= \Phi(\psi_0 + \psi_1 x_i + \psi_2 z_i + \psi_3 y_i),\end{aligned}$$

where Φ is the probit function (as before), y_i is an indicator for HIV status (1 or 0), x_i includes demographic and socioeconomic characteristics (sex, age, age squared, marital status, education level, employment status, household size, number of children, log per capita consumption, and asset-based wealth), variables that capture individual risk of HIV infection (biomarkers and HIV knowledge), stigmatizing attitudes, and neighborhood dummy variables; (β_0, β_1) and

$(\psi_0, \psi_1, \psi_2, \psi_3)$ are distinct; the vector z_i consists of the identity codes of the nurses who administered the HIV test in the biomedical survey. These variables were not included in the equation for HIV, because these nurses can influence the missingness rate (some of them are more persuasive than others), but they cannot influence the outcome of the test directly, because they are randomly assigned to households. That is, the random assignment of interviewers justifies the exclusion of z_i from the model for y_i , and allows the model parameters to be estimated consistently.

The authors found that the bias resulting from refusal was not significant for the overall sample. However, a detailed analysis using kernel density estimates indicated that the bias is substantial for the younger and the poorer populations. Nonparticipants in these subsamples are estimated to be three times more likely to be HIV-positive than participants. The difference is particularly pronounced for women. The authors argue that estimates of prevalence rates that ignore this selection effect may be seriously biased for specific target groups, leading to misallocation of resources for prevention and treatment.

Example 15.13 A Bivariate Normal Pattern-Mixture Model with Parameter Restrictions. For the pattern-mixture model in Example 15.8, write $Y = Y_2$ and $X = Y_1$, so that the data consist of a random sample of n units (y_{i1}, y_{i2}) , $i = 1, \dots, n$ on (Y_1, Y_2) , with $\{y_{i1}\}$ fully observed and $\{y_{i2}\}$ observed for $i = 1, \dots, r$ and missing for $i = r+1, \dots, n$. Consider the following normal pattern-mixture model for the joint distribution of (Y_1, Y_2) :

$$(y_{i1}, y_{i2} | m_i = m, \phi) \sim_{\text{iid}} N(\mu^{(m)}, \Sigma^{(m)}), \quad m_i \sim_{\text{iid}} \text{BERN}(1 - \pi), \quad m = 0, 1, \quad (15.15)$$

where π is the response rate, which implies the model (15.12) for the conditional distribution of Y_2 given Y_1 . This model has 11 parameters $\phi = (\pi, \phi^{(0)}, \phi^{(1)})$ where $\phi^{(m)} = (\mu_1^{(m)}, \mu_2^{(m)}, \sigma_{11}^{(m)}, \sigma_{22}^{(m)}, \sigma_{12}^{(m)})$, $m = 0, 1$. The likelihood is

$$\begin{aligned} L(\phi_{\text{id}} | Y_{(0)}) &= \pi^r (1 - \pi)^{n-r} \prod_{i=1}^r f(y_{i1}, y_{i2} | m_i = 0, \phi^{(0)}) \\ &\times \prod_{i=r+1}^n f(y_{i1} | m_i = 1, \mu^{(1)}, \sigma_{11}^{(1)}) \end{aligned} \quad (15.16)$$

where ϕ_{id} is the subset of eight of the parameters that appears in the likelihood, namely

$$\phi_{\text{id}} = \left(\pi, \mu_1^{(0)}, \mu_2^{(0)}, \sigma_{11}^{(0)}, \sigma_{12}^{(0)}, \sigma_{22}^{(0)}, \mu_1^{(1)}, \sigma_{11}^{(1)} \right) \quad (15.17)$$

The remaining parameters in ϕ , namely $\phi_{\text{nid}} = (\mu_2^{(1)}, \sigma_{12}^{(1)}, \sigma_{22}^{(1)})$, do not appear in the likelihood, but can be estimated by restrictions based on the assumed

nature of the missingness mechanism. First, suppose that missingness of Y_2 depends only on Y_1 ; that is, the mechanism is MAR. Then the distribution of Y_2 given Y_1 is the same for both missingness patterns (as seen in Example 1.13), implying the following restrictions on the parameters of this distribution:

$$\beta_{20 \cdot 1}^{(0)} = \beta_{20 \cdot 1}^{(1)}, \beta_{21 \cdot 1}^{(0)} = \beta_{21 \cdot 1}^{(1)}, \sigma_{22 \cdot 1}^{(0)} = \sigma_{22 \cdot 1}^{(1)} \quad (15.18)$$

The three linear restrictions in (15.18) equals the number of inestimable parameters in ϕ_{nid} . As a result, the ML estimates $\hat{\phi}_{\text{id}}$ of ϕ_{id} are unconstrained, and they are

$$\hat{\pi} = r/n, \mu_j^{(0)} = \bar{y}_j, \hat{\sigma}_{jk}^{(0)} = s_{jk}, \hat{\mu}_1^{(1)} = \bar{y}_1^{(1)}, \hat{\sigma}_{11}^{(1)} = s_{11}^{(1)} \quad (15.19)$$

respectively, the sample proportion of incomplete units, the sample mean and covariance matrix for the complete units, and the mean and variance of Y_1 for the incomplete units. In a Bayesian analysis, draws $\phi_{\text{id}}^{(d)}$ of ϕ_{id} are straightforward under a beta prior distribution for π , and the inverse chi-squared/normal prior distributions for $\phi^{(0)}$ and $\phi^{(1)}$ (see Section 7.3 for details).

ML estimates of other parameters are obtained by expressing them as functions of ϕ_{id} and replacing ϕ_{id} by its ML estimate $\hat{\phi}_{\text{id}}$; draws from their posterior distributions are obtained by replacing ϕ_{id} by draws $\phi_{\text{id}}^{(d)}$. In particular, Eq. (15.18) implies that

$$\mu_2^{(1)} = \beta_{20 \cdot 1}^{(1)} + \beta_{21 \cdot 1}^{(1)} \mu_1^{(1)} = \beta_{20 \cdot 1}^{(0)} + \beta_{21 \cdot 1}^{(0)} \mu_1^{(1)}$$

and the overall mean of Y_2 is

$$\mu_2 = \pi \mu_2^{(0)} + (1 - \pi) \mu_2^{(1)} = \pi \mu_2^{(0)} + (1 - \pi) \left(\beta_{20 \cdot 1}^{(0)} + \beta_{21 \cdot 1}^{(0)} \mu_1^{(1)} \right);$$

hence the ML estimate of $\mu_2^{(1)}$ is $\hat{\beta}_{20 \cdot 1}^{(0)} + \hat{\beta}_{21 \cdot 1}^{(0)} \hat{\mu}_1^{(1)} = \bar{y}_2 + b_{21 \cdot 1} (\bar{y}_1^{(1)} - \bar{y}_1)$, where $b_{21 \cdot 1} = (s_{12}/s_{11})$ is the slope of the regression of Y_2 on Y_1 estimated on the complete units. The ML estimate of μ_2 is

$$\hat{\mu}_2 = \pi \hat{\mu}_2^{(0)} + (1 - \pi) \left(\hat{\beta}_{20 \cdot 1}^{(0)} + \hat{\beta}_{21 \cdot 1}^{(0)} \mu_1^{(1)} \right) = \bar{y}_2 + \frac{n-r}{n} b_{21 \cdot 1} (\bar{y}_1^{(1)} - \bar{y}_1)$$

which is the regression estimate Eq. (7.9), and is also the ML estimate of μ_2 for the ignorable selection model. ML estimates of the other parameters of the marginal distribution of Y_1 and Y_2 are also the same as those for the ignorable selection model. Hence, the pattern-mixture model (15.15) with restrictions (15.18) and the ignorable normal selection model of Section 7.2.1 together yield the same ML estimates, despite differing distributional assumptions.

Now suppose that missingness of Y_2 is assumed to depend on Y_2 but not Y_1 . This assumption implies that the distribution of Y_1 given Y_2 and M does not

depend on M , and hence the parameters of the regression of Y_1 given Y_2 are the same for each missingness pattern, that is

$$\beta_{10 \cdot 2}^{(0)} = \beta_{10 \cdot 2}^{(1)}, \beta_{12 \cdot 2}^{(0)} = \beta_{12 \cdot 2}^{(1)}, \sigma_{11 \cdot 2}^{(0)} = \sigma_{11 \cdot 2}^{(1)}. \quad (15.20)$$

These three restrictions again yield ML estimates of ϕ_{id} that are the same as for the MAR model (subject to one caveat, which is mentioned below). We can write the mean of Y_1 for incomplete units in terms of ϕ_{id} as follows:

$$\mu_2^{(1)} = \left(\mu_1^{(1)} - \beta_{10 \cdot 2}^{(1)} \right) / \beta_{12 \cdot 2}^{(1)} = \left(\mu_1^{(1)} - \beta_{10 \cdot 2}^{(0)} \right) / \beta_{12 \cdot 2}^{(0)} \quad (15.21)$$

Substituting ML estimates of ϕ_{id} yields $\hat{\mu}_2^{(1)} = \bar{y}_2 + (\bar{y}_1^{(1)} - \bar{y}_1) / b_{12 \cdot 2}$, where $b_{12 \cdot 2} = s_{12} / s_{22}$ is the slope of the regression of Y_1 on Y_2 estimated on the complete units. Hence

$$\hat{\mu}_2 = \bar{y}_2 + (1/b_{12 \cdot 2})(\hat{\mu}_1 - \bar{y}_1), \quad (15.22)$$

an estimator of the mean of Y_2 first proposed by Brown (1990), we believe; Eq. (15.22) effectively imputes missing values of Y_2 using the inverse regression of Y_1 on Y_2 , as in calibration problems. Similar arguments yield the following ML estimates of the other parameters:

$$\hat{\sigma}_{12} = s_{12} + (1/b_{12 \cdot 2})(\hat{\sigma}_{11} - s_{11}) \quad (15.23)$$

and

$$\hat{\sigma}_{22} = s_{22} + (1/b_{12 \cdot 2})^2(\hat{\sigma}_{11} - s_{11}). \quad (15.24)$$

The caveat is that the restrictions (15.20), unlike the MAR restrictions (15.18), involve parameters that are not distinct from the identified parameters ϕ_{id} . As a result, the ML estimates, or draws from the posterior distribution of ϕ_{id} , may need modifications to ensure that parameters lie within their respective parameter spaces. In particular if $\hat{\sigma}_{11 \cdot 2}^{(0)} > \hat{\sigma}_{11}^{(1)}$, then $\hat{\sigma}_{22}^{(1)}$ is negative, and cannot be the ML estimate of $\sigma_{22}^{(1)}$. In that case, ML estimates of $\sigma_{12}^{(1)}$ and $\sigma_{22}^{(1)}$ are set to zero, and $\sigma_{11 \cdot 2}^{(0)}$ and $\sigma_{11}^{(1)}$ are both estimated by pooling the residual variance of Y_1 given Y_2 for complete units with the variance of Y_1 for incomplete units.

The large-sample sampling variances of (15.22)–(15.24) are given by Taylor Series calculations (Little 1994). A better approach for small samples is to incorporate a prior distribution for the parameters and simulate draws from their joint posterior distribution. With Jeffreys' prior distributions, draws of ϕ_{id} from their posterior distribution can be obtained via the following eight steps:

- (i) $\pi \sim \beta(n - r + 0.5, r + 0.5)$
- (ii) $1/\sigma_{22}^{(0)} \sim \chi_{r-1}^2 / (rs_{22})$

- (iii) $1/\sigma_{11}^{(1)} \sim \chi^2_{n-r-1}/((n-r)s_{11}^{(1)})$
- (iv) $1/\sigma_{11\cdot 2}^{(0)} \sim \chi^2_{r-2}/(rs_{11\cdot 2})$
- (v) $\beta_{12\cdot 2}^{(0)} \sim N(b_{12\cdot 2}, \sigma_{11\cdot 2}^{(0)}/(rs_{22}))$
- (vi) $\beta_{10\cdot 2}^{(0)} \sim N(\bar{y}_1 - \beta_{12\cdot 2}^{(0)}\bar{y}_2, \sigma_{11\cdot 2}^{(0)}/r)$
- (vii) $\mu_2^{(0)} \sim N(\bar{y}_2, \sigma_{22}^{(0)}/r)$
- (viii) $\mu_1^{(1)} \sim N(\bar{y}_1^{(1)}, \sigma_{11}^{(1)}/(n-r))$.

To satisfy the parameter constraints, the drawn value of $\sigma_{11}^{(1)}$ from (iii) must be greater than the drawn value of $\sigma_{11\cdot 2}^{(0)}$ from (iv); if this is not the case, then these draws are discarded and steps (iii) and (iv) are repeated. Draws from the posterior distributions of other parameters are obtained by expressing them as functions of ϕ_{id} and then substituting the drawn value of ϕ_{id} .

A feature of this pattern-mixture model is that a parametric model is not required for the missingness mechanism, assumptions about the mechanism being conveyed through parameter restrictions. For selection models, such a parametric model is not needed under MAR, but is needed for MNAR models where missingness depends on Y_2 . On the other hand, the model does require joint normality of Y_1 and Y_2 within each missingness pattern, often not an innocuous assumption.

Example 15.14 Nonresponse Adjustment of Survey Estimates Based on Auxiliary Variables Subject to Measurement Error. West and Little (2013) model MNAR data using a multivariate extension of the MNAR model in Example 15.13. Suppose that $Y = (Y_1, Y_2, Y_3)$, where (i) Y_2 and Y_3 are survey variables subject to nonresponse, observed for $i = 1, \dots, r$ and missing for $i = r+1, \dots, n$, and (ii) Y_1 is a proxy variable for Y_2 that is fully observed for $i = 1, \dots, n$. As an example, suppose that Y_2 is income, and Y_1 is a fully observed variable related to income, such as an estimate of the value of a house of a sampled individual, available from external sources. Missingness of Y_2 and Y_3 is assumed to be related to Y_2 but not Y_1 or Y_3 , given Y_2 , the logic being that missingness is related to the true value (Y_2) rather than the proxy (Y_1), as would be implied by the MAR assumption. Parameters of the distribution of Y_2 and Y_3 are estimated under the following extension of the model in Example 15.13:

$$(y_{i1}, y_{i2}, y_{i3} | m_i = m, \phi) \sim_{iid} N(\mu^{(m)}, \Sigma^{(m)}), \quad m_i \sim_{iid} \text{BERN}(1 - \pi), \quad m = 0, 1, \quad (15.25)$$

with parameters restrictions under the assumption that missingness of Y_2 and Y_3 depends on $Y = (Y_1, Y_2, Y_3)$ only through Y_2 , which implies that the parameters of the regression of Y_1 and Y_3 on Y_2 are the same for complete and incomplete units. The number of restrictions equals the number of estimable parameters, and ML and Bayes estimates are straightforward extensions of the estimates for the corresponding MNAR model in Example 15.13.

Validity of the resulting estimates depend crucially on the assumption that missingness depends only on Y_2 , which is arguably more plausible than MAR if Y_1 differs from Y_2 by a random measurement error. West and Little (2013) discuss extensions of the model (15.25) to include fully observed covariates Z , with missingness depending on Y_2 and Z .

15.3.5 Sensitivity Analysis

In situations where following up a subsample of nonrespondents is not feasible, and the assumptions required to identify a specific MNAR model are not tenable, another approach is to conduct a sensitivity analysis to assess the effect of deviations from MAR. The next example illustrates this approach with a generalization of the pattern-mixture model of Example 15.13.

Example 15.15 *Sensitivity Analysis for the Bivariate Normal Pattern-Mixture Model.* For the model in Eq. (15.15), suppose that missingness of Y_2 given Y_1 and Y_2 is assumed to depend only on $Y_2^* = Y_1 + \lambda Y_2$, and assume for the present that λ is known. Then the conditional distribution of Y_1 given Y_2^* is independent of missingness pattern, that is, for unit i :

$$f(y_{i1}|y_{i2}^*, \phi, m_i = 1) = f(y_{i1}|y_{i2}^*, \phi, m_i = 0) \quad (15.26)$$

ML estimates of the mean and variance of Y_2^* and the covariance of Y_1 and Y_2^* , then have the form discussed in Example 15.13. Transformations yield the following ML estimates of μ and Σ :

$$\hat{\mu}_2 = \bar{y}_2 + b_{21 \cdot 1}^{(\lambda)} (\hat{\mu}_1 - \bar{y}_1), \quad (15.27)$$

$$\hat{\sigma}_{22} = s_{22} + (b_{21 \cdot 1}^{(\lambda)})^2 (\hat{\sigma}_{11} - s_{11}), \quad (15.28)$$

and

$$\hat{\sigma}_{12} = s_{12} + b_{21 \cdot 1}^{(\lambda)} (\hat{\sigma}_{11} - s_{11}), \quad (15.29)$$

where

$$b_{21 \cdot 1}^{(\lambda)} = \frac{\lambda s_{22} + s_{12}}{\lambda s_{12} + s_{11}} \quad (15.30)$$

These expressions yield the ignorable ML (IML) estimates of Section 7.2.1 when $\lambda = 0$, and the ML estimates of Example 15.13 in the limit as $\lambda \rightarrow \infty$. Negative values of λ are also possible; for example, if missingness depends on the change $Y_2 - Y_1$, then $\lambda = -1$. It can be shown (Problems 15.10 and 15.11)

that when $\lambda = -\beta_{12-2}^{(0)}$, the complete-case estimate \bar{y}_2 is the ML estimate of the mean of Y_2 , and the available-case estimate $\hat{\mu}_1 - \bar{y}_2$ is the ML estimate of the difference in means. For a Bayesian analysis, draws of parameters of the joint distribution of Y_1 and Y_2 replace the ML estimates, using methods analogous to those in Example 15.13.

As with MNAR selection models, the data supply no evidence for λ : the fit of the model to the observed data is identical for all choices of λ , provided estimates lie within their respective parameter spaces. This limitation arises because there are no data for estimating the distribution of Y_2 given Y_1 for the incomplete units. Uncertainty about the choice of λ can be reflected in the inference by specifying a prior distribution; alternatively inferences about the parameters can be displayed for a range of plausible values of λ , to assess sensitivity of inferences to the missingness mechanism, as illustrated in the next example.

Figure 15.1 displays 95% intervals for the mean of Y_2 for two artificial data sets, with statistics $(\bar{y}_1, \bar{y}_2, s_{11}, s_{12}, s_{22}, \bar{y}_1^{(1)}, s_{11}^{(1)})$ generated according to the normal pattern-mixture model with $\pi = 1/3$, $\mu_1^{(0)} = \mu_2^{(0)} = 0$, $\mu_1^{(1)} = 1$, $\sigma_{11}^{(0)} = \sigma_{11}^{(1)} = \sigma_{22}^{(0)} = 1$, $\sigma_{12}^{(0)} = \rho^{(0)} = 0.4$ or 0.8 . One data set was created for each choice of $\rho^{(0)}$, with sample sizes $n = 75$, $r = 50$, $n - r = 25$.

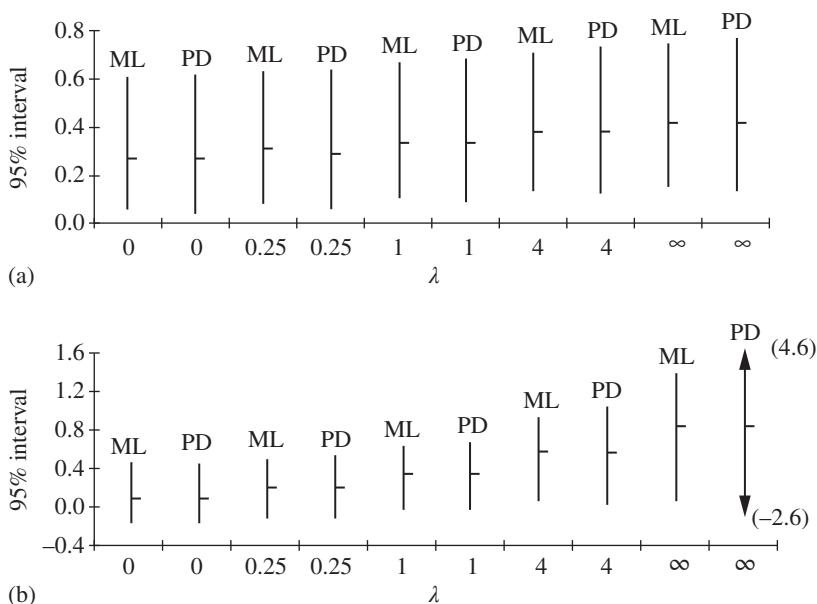


Figure 15.1 Example 15.15, 95% intervals for generated data with (a) $\rho = 0.8$ and (b) $\rho = 0.4$.

For each value of λ , two 95% intervals are displayed:

- (i) The ML estimate ± 2 asymptotic standard errors (labeled ML); and
- (ii) the posterior mean ± 2 posterior standard deviations, computed from 5000 draws from the posterior distribution of μ_2 with Jeffreys' prior distributions (labeled PD).

The mark in each interval is the true value of μ_2 assuming that the chosen value of λ is in fact correct. This quantity is computed via a population analog of (15.23):

$$\mu_2 = \mu_2^{(0)} + \pi \left(\frac{\sigma_{12} + \lambda\sigma_{22}}{\sigma_{11} + \lambda\sigma_{12}} \right) (\mu_1^{(1)} - \mu_1^{(0)}).$$

The positive values of μ_2 reflect the fact that Y_1 and Y_2 are positively correlated and $\mu_1^{(1)}$ is greater than $\mu_1^{(0)}$; μ_2 increases with λ . The sample intervals for μ_2 also shift upward with λ , and all cover the true mean provided the true value of λ is chosen. The intervals become wider with increasing λ , reflecting increasing uncertainty. The locations and widths of the intervals are much more sensitive to λ for the data set with $\rho = 0.4$ than the data set with $\rho = 0.8$, illustrating the utility of a highly correlated fully observed predictor Y_1 .

The Bayes intervals are always wider than the ML intervals and provide a better reflection of uncertainty than the ML intervals. The most extreme difference is when $\lambda = \infty$ and $\rho = 0.4$, when the Bayesian interval $(-2.6, 4.6)$ is extreme and off the chart, reflecting problems when the drawn values of $\beta_{12.2}$ approach zero, because these draws appear in the denominator of the adjustment of $\mu_2^{(0)}$. The $\lambda = \infty$ model requires a strong correlation or large sample size to keep the estimates of $\beta_{12.2}$ away from zero and hence, allow reliable estimates of μ_2 .

Little and Wang (1996) extend the bivariate normal pattern-mixture model (15.15) with parameter restrictions to multivariate Y_1 , Y_2 and covariates X , with Y_1 and X fully observed and Y_2 observed for r units and missing for $n - r$ units. Extensions of the model to more than two patterns of missing data appear problematic (Tang et al. 2003).

Andridge and Little (2011) propose an extension of (15.15) to (Y_2, Z) , where Y_2 has missing values and Z is a set of fully observed auxiliary variables. They propose “proxy pattern-mixture analysis” (PPMA), where the model (15.15) applied with Y_1 is replaced by Y_1^* , the best predictor of Y_2 based on a regression of Y_2 on Z , estimated from the units, where both Y_2 and Z are observed. When Y_1^* is scaled to have the same variance as Y_2 , it is called the “best proxy” for Y_2 based on the auxiliary data Z . This approach provides a sensitivity analysis for deviations from MAR for unit nonresponse in surveys, where Y_2 is a survey variable and Z is the set of variables observed for respondents and nonrespondents. Andridge and Little (2009) extends PPMA to a binary Y_2 .

Pattern	Unit, i	z_i	w_i	x_i	y_i
1	$i = 1, \dots, m$	✓	✓	✓	✗
2	$i = m + 1, \dots, m + r$	✓	✓	✗	✗
3	$i = m + r + 1, \dots, n$	✓	✗	✓	✗

Key: ✓ denotes observed, ✗ denotes observed or missing some components.

Figure 15.2 Missing data structure for subsample ignorable likelihood methods.

15.3.6 Subsample Ignorable Likelihood for Regression with Missing Data

Little and Zhang (2011) consider data with the missing data pattern of Figure 15.2, with four sets of variables (Z, W, X, Y), all of which can be vectors. Interest concerns the regression of Y on (Z, W, X) , specifically the parameters ϕ of the distribution of Y given (Z, W, X) , say $p(y_i|z_i, w_i, x_i, \phi)$. The covariates (Z, W, X) are partitioned into three sets: Z includes any fully observed covariates, whereas covariates W and X have missing values and are distinguished by different assumptions about their missingness mechanisms. Specifically, the following assumptions are made concerning the missingness of (W, X, Y) :

- (a) *Covariate missingness of W* : The probability that W is fully observed depends only on the covariates and not Y . Specifically, let m_{w_i} denote the vector of missingness indicators for w_i , with entries 0 for variables that are observed and 1 for variables that are missing. Write $m_{w_i} = \underline{0}$ if all the entries in m_{w_i} are zero, so that w_i is fully observed. Then it is assumed that

$$\Pr(m_{w_i} = \underline{0}|z_i, w_i, x_i, y_i, \psi_w) = \Pr(m_{w_i} = \underline{0}|z_i, w_i, x_i, \psi_w) \quad \text{for all } y_i. \quad (15.31)$$

- (b) *Subsample MAR of X and Y , given that W is observed*: Missingness of X and Y is MAR within the subsample of units i for which w_i is fully observed, that is, for which $m_{w_i} = \underline{0}$. Specifically, let $m_{(x_i, y_i)}$ denote the vector of missingness indicators for (x_i, y_i) . Then it is assumed that

$$\begin{aligned} & \Pr(m_{(x_i, y_i)}|z_i, w_i, x_i, y_i, m_{w_i} = \underline{0}, \psi_{xy,w}) \\ &= \Pr(m_{(x_i, y_i)}|z_i, w_i, x_{(0)i}, y_{(0)i}, m_{w_i} = \underline{0}, \psi_{xy,w}) \quad \text{for all } x_{(1)i}, y_{(1)i}, \end{aligned} \quad (15.32)$$

where $(x_{(0)i}, y_{(0)i})$ and $(x_{(1)i}, y_{(1)i})$ represent the observed and missing components of (x_i, y_i) , respectively. Note that Eqs. (15.31) and (15.32) are in general MNAR assumptions, because missingness of w_i can depend on missing values of w_i and x_i , and missingness of (x_i, y_i) can depend on missing values of w_i .

A “subsample ignorable likelihood” (SSIL) method applies a likelihood method (that is, ML or Bayes), ignoring the missingness mechanism for the subsample of units for which W is fully observed, that is discarding the units where any components of W are missing. Under the assumptions (15.31) and (15.32),

Little and Zhang (2011) show that the data are partially MAR, in the sense of Definition 6.5, for direct-likelihood inferences for ϕ (the parameters of the conditional distribution of Y given Z , W and X). Hence, a SSIL method is valid in the “partial likelihood” sense of (Cox 1975), although it is not a full likelihood method and hence, may not be fully efficient. Intuitively, covariate missingness of W justifies the restriction to units with W observed, and subsample MAR of X and Y allows the missingness mechanism for these variables to be ignored in the subset of units with W observed.

Example 15.16 Application to Regression of Blood Pressure. To illustrate SSIL, Little and Zhang (2011) apply it to the US National Health and Nutrition Examination Survey (NHANES, 2003 and 2004) data to study the relationship of socioeconomic variables and blood pressure. They regress the two blood pressure measurements, systolic blood pressure (SBP) and diastolic blood pressure (DBP), on household income (HHINC, in 1000 \$/yr), years of education (EDU, in years), age (in years), sex, and body mass index (BMI, kg/m²). HHINC data are categorical with 11 categories in the NHANES, and they use the median HHINC within each category as a proxy for true HHINC.

HHINC, education, BMI, and the two blood pressure measures are subject to missingness as indicated in Table 15.7. We allow MNAR for HHINC, because people with high or low income are more likely to fail to report it due to privacy concerns, whereas we assume MAR missingness for other variables. For illustrative purposes, we ignore the design features (weighting and clustering, etc.) of the NHANES study. The results of complete-case (CC) analysis, ignorable ML (IML) analysis, and subsample ignorable maximum likelihood (SSIML) are shown in Table 15.8. CC and SSIML yield similar results, but the SSIML gives smaller standard error estimates, reflecting a possible gain in efficiency.

Table 15.7 Example 15.16, percentage of missingness in NHANES 2003–2004

	Full data (<i>n</i> = 9041)	Subset with HHINC observed (<i>n</i> = 5400)
Household income (%)	40.27	0.00
Education (%)	17.24	16.74
Age (%)	0.00	0.00
Gender (%)	0.00	0.00
BMI (%)	9.84	9.48
SBP/DBP (%)	25.02	24.50

Source: Little and Zhang (2011). Reproduced with permission of John Wiley and Sons.

Table 15.8 Example 15.16, estimates of linear regression of blood pressure on measures of social inequality (NHANES 2003–2004)

	CC analysis			IML analysis			SSMLM analysis		
	Estimate	SE	p-Value	Estimate	SE	p-Value	Estimate	SE	p-Value
<i>Systolic blood pressure (SBP)</i>									
Intercept	87.80	1.16	<0.0001	89.28	1.06	<0.0001	87.53	1.35	<0.0001
HHINC	-0.01	0.01	0.3907	-0.01	0.01	0.4574	-0.01	0.01	0.3482
EDU	-2.30	0.57	<0.0001	-2.06	0.44	<0.0001	-2.38	0.55	<0.0001
Age	0.49	0.01	<0.0001	0.50	0.01	<0.0001	0.50	0.01	<0.0001
Female	3.31	0.48	<0.0001	2.78	0.44	<0.0001	3.15	0.46	<0.0001
BMI	0.46	0.04	<0.0001	0.41	0.03	<0.0001	0.47	0.04	<0.0001
<i>Diastolic blood pressure (DBP)</i>									
Intercept	45.46	1.06	<0.0001	46.94	1.00	<0.0001	45.46	1.19	<0.0001
HHINC	0.03	0.01	0.0008	0.03	0.01	0.0026	0.03	0.01	0.005
EDU	4.86	0.52	<0.0001	4.06	0.43	<0.0001	4.95	0.52	<0.0001
Age	0.12	0.01	<0.0001	0.11	0.01	<0.0001	0.11	0.01	<0.0001
Female	1.81	0.44	<0.0001	1.83	0.36	<0.0001	1.86	0.42	<0.0001
BMI	0.43	0.04	<0.0001	0.40	0.03	<0.0001	0.44	0.04	<0.0001

Source: Little and Zhang (2011). Reproduced with permission of John Wiley and Sons.

SSIL analysis has the following strengths: (i) it is simple and easy to implement, because software for doing ignorable likelihood analyses is all that is required, and this software is now widely available for many models. (ii) It avoids discarding all incomplete units. (iii) The method yields consistent estimates for the missingness mechanisms defined by (15.31) and (15.32), under which both ignorable likelihood methods and complete-case analysis fail to give consistent estimates.

In practice, the main challenge in applying SSIL is deciding which covariates belong in the set W and which belong in the set X , that is, which covariates are used to create the subsample for the MAR analysis. The choice is guided by the basic assumptions, (15.31) and (15.32), concerning which variables are considered covariate-dependent MNAR and which are considered subsample MAR. This is a substantive choice that requires understanding the missingness mechanism in the particular context. It is aided by learning more about the missingness mechanism, as from recording reasons why particular values are missing. All missing-data methods make assumptions about the missingness mechanism, and these need to be as reasonable and well considered as possible.

Von Hippel (2007) applies MAR-based MI in the regression setting where both predictors and outcome Y have missing values, and then applies the final regression analysis to the subsample of units with Y observed, that is dropping the units with Y imputed. This strategy reduces the simulation error from MI, and it can be viewed as a special case of SSIL.

15.4 Other Models and Methods for MNAR Missing Data

15.4.1 MNAR Models for Repeated-Measures Data

Let y_i denote a vector of repeated measures, and x_i a set of fixed covariates. Mixed-effect models for y_i given x_i that modeled the repeated measures using unobserved within-subject random effects β_i were discussed in Section 11.5, assuming MAR. With MNAR missingness, a variety of models can be based on various factorizations of the joint distribution of $y_i m_i$, and β_i given x_i . Three particular factorizations may make substantive sense (Little 2008):

Mixed-effect selection models of the form

$$f(y_i, m_i, \beta_i | x_i, \gamma_1, \gamma_2, \phi) = f(\beta_i | x_i, \gamma_1) f(y_i | x_i, \beta_i, \gamma_2) f(m_i | x_i, y_i, \beta_i, \phi); \quad (15.33)$$

Mixed-effect pattern-mixture models of the form

$$f(y_i, m_i, \beta_i | x_i, \gamma_1, \gamma_2, \phi) = f(m_i | x_i, \phi) f(\beta_i | x_i, m_i, \gamma) f(y_i | x_i, m_i, \beta_i, \gamma_2); \quad (15.34)$$

Mixed-effect hybrid models of the form

$$f(y_i, m_i, \beta_i | x_i, \gamma_1, \gamma_2, \phi) = f(\beta_i | x_i, \gamma_1) f(m_i | x_i, \beta_i, \phi) f(y_i | x_i, m_i, \beta_i, \gamma_2). \quad (15.35)$$

Examples in the literature do not always make a compelling argument for the choice of missingness mechanism in the particular applied setting, but we view it as a key to success. Also, the lack of parameter estimability remains a problem, suggesting that a sensitivity analysis may be preferable to assuming unjustified parameter restrictions.

Little (1995, 2008) discusses various special cases of Eqs. (15.33)–(15.35):

1. *MNAR outcome-dependent dropout*: where dropout depends on missing components of y_i , such as the (unrecorded) value of the outcome at the time when the subject drops out, but not on the random effects β_i . Under this assumption, the factor on the right side of Eq. (15.33) has the form

$$f(m_i|x_i, y_i, \beta_i, \phi) = f(m_i|x_i, y_i, \phi). \quad (15.36)$$

Diggle and Kenward (1994) assumes this drop-out mechanism to analyze data from a longitudinal milk protein trial. Cows were randomly allocated to one of three diets (barley, mixed barley and lupins, and lupins), and the protein content of milk samples taken weekly for a period of 20 weeks was assayed. “Dropout” corresponded to cows that stopped producing milk before the end of the experiment. The complete-data model $f_Y(y_i|x_i, \beta_i)$ specified a quadratic model for the mean protein content over time, with an intercept that depended on diet (thus modeling an additive effect of treatment). The covariance structure was assumed to be a combination of an autoregressive structure with an added independent measurement error. The dropout distribution $f_M(m_i|x_i, y_i, \phi)$ was modeled as depending on the current and previous value of protein content, specifically:

$$\text{logit}\{\Pr(m_{it} = 1|x_i, m_{i,t-1} = 0, y_i, \phi)\} = \phi_0 t + \phi_1 y_{i,t-1} + \phi_2 y_{it}.$$

ML estimates of the coefficients, $\phi_1 = 12.0$, $\phi_2 = -20.4$, suggested that the probability of “dropout” increases when the prevailing level of protein is low or the increment between the last and current protein contents is high.

We see two problems with this analysis. First, we do not think that “dropout” meets our general definition of missing data, in that it does not seem meaningful to consider what the protein content would have been, had a cow’s milk not dried up. Second, considering concerns about assumptions needed to make (ϕ_1, ϕ_2) estimable, a better approach might be to conduct a sensitivity analysis for a range of plausible alternative choices of (ϕ_1, ϕ_2) .

2. *MNAR random-coefficient dependent dropout*: Another form of MNAR dropout model assumes dropout at time t depends on the value of β_i , that is

$$f_M(m_i|x_i, y_i, \beta_i, \phi) = f_M(m_i|x_i \beta_i, \phi) \quad (15.37)$$

Examples of models of dropout of the form (15.37) include Wu and Carroll (1988), Shih et al. (1994), Mori et al. (1994), Schluchter (1992), and DeGruttola and Tu (1994), who modeled the relationship between the progression of CD4 lymphocyte count and survival for patients enrolled in a clinical trial of two alternative doses of zidovudine. As with the previous example, a problem with their approach is that the selection model factorization effectively treats the CD4 counts after death as missing values, which is not in accord with our definition of missing data. A more appropriate analysis would condition the analysis of CD4 counts at any time on individuals who have survived up to that time, using principal stratification methods (Frangakis and Rubin 2002).

3. *Shared parameter models:* These models assume both the outcome process and the dropout process depend on shared latent variables. Examples include Ten Have et al. (1998, 2002), Albert et al. (2002), and Roy (2003). They are special cases of (15.33) and (15.35), where y_i and m_i are assumed independent given β_i :

$$f(y_i, m_i, \beta_i | x_i, \gamma_1, \gamma_2, \phi) = f(\beta_i | x_i, \gamma_1) f(y_i | x_i, \beta_i, \gamma_2) f(m_i | x_i, \beta_i, \phi). \quad (15.38)$$

Albert et al. (2002) analyze data from a clinical trial of treatments of heroin addiction that randomized patients into one of two treatment groups, buprenorphine ($n = 53$) and methadone ($n = 55$). Patients were scheduled for urine tests three times a week for 17 weeks postrandomization (51 scheduled responses). The outcome y_{it} at time t was a binary variable for the presence or absence of opiates at each follow-up visit.

The analysis was complicated by unequally spaced visits and the large amount of missing data, which took the form of dropouts and intermittent missing data. A number of subjects withdrew from the study, because of poor compliance, or because they were offered places in treatment programs that gave unmasked treatment and long-term care. Intermittent missingness was thought to be more closely associated with the outcomes, because patients may be less likely to show up when they are taking opiates. The proportion of patients dropping out by the end of the 17-week period was 80% in the methadone group and 59% in the buprenorphine group. In addition, patients had a sizable amount of intermittent missing data, with a higher proportion in the buprenorphine arm than the methadone arm. Also, the Spearman rank correlation between the proportion of positive tests and the time to dropout was -0.44 in the buprenorphine arm and -0.10 in the methadone arm. The correlations between the proportion of positive tests and the proportion of intermittent missing visits before dropout in the buprenorphine and methadone arms were 0.40 and 0.29 , respectively. These calculations suggest that addicts who are more likely to use drugs are both more likely to dropout and to have a higher frequency of intermittent missing data.

before dropout than addicts who use opiates less frequently. The shared parameter model (15.38) assumes that missingness is related to the underlying level and trend in the presence of opiates; an alternative approach is to model the association of missingness at time t with the presence or absence of a positive test at time t , as in Eq. (15.36).

MNAR models for repeated-measures data based on the pattern-mixture and hybrid factorizations (15.34) and (15.35) are discussed in Little (1995, 2008) and in Yuan and Little (2009). Details are omitted here.

15.4.2 MNAR Models for Categorical Data

At least two types of MNAR models for incomplete categorical data have been considered. Pregibon (1977), Little (1982), and Nordheim (1984) introduce prior odds of response for categories of the contingency table that modify the likelihood. Hierarchical loglinear models for the joint distribution of the categorical variables and indicator variables for nonresponse are considered by Baker and Laird (1988), Fay (1986), and Little (1985b). We consider the latter approach here, as it is closer in spirit to the contingency table models discussed in Chapter 13. Unlike those models, the MNAR models discussed here involve subtle issues of estimability, which we do not address in detail here. Attention is confined to a two-way contingency table with one supplemental margin, to convey basic ideas.

Example 15.17 Two-Way Contingency Table with One Supplemental Margin. Suppose data are as in Example 13.1, with n units with two categorical variables, Y_1 with levels $j = 1, \dots, J$ and Y_2 with levels $k = 1, \dots, K$. r completely classified units that form a two-way contingency table $\{r_{jk}\}$ and $m = n - r$ units classified by Y_1 but not by Y_2 that form a supplemental margin $\{m_j\}$. For illustration, we fit models to the data set in Table 15.9, with $J = K = 2$.

Define M to take the value 1 if Y_2 is missing, 0 if Y_2 is observed. Suppose that for fixed n , complete units have a multinomial distribution over the

Table 15.9 Example 15.17, a 2×2 contingency table with one partially classified margin

		Y_2				Y_2			
		1	2			1	2		
Y_1	1	$r_{11} = 100$	$r_{12} = 20$	$r_{1+} = 120$	Y_1	1	$m_{11} = ?$	$m_{12} = ?$	$m_1 = 40$
	2	$r_{21} = 30$	$r_{22} = 50$	$r_{2+} = 80$		2	$m_{21} = ?$	$m_{22} = ?$	$M_2 = 60$
		$r_{+1} = 130$	$r_{+2} = 70$	$r = 200$					$m = 100$
Fully classified ($M = 0$)				Partially classified ($M = 1$)					

$J \times K \times 2$ table formed by Y_1 , Y_2 , and M . Let $\pi_{jk} = \Pr(Y_1 = j, Y_2 = k)$ and $\phi_{jk} = \Pr(M = 1 | Y_1 = j, Y_2 = k)$ so that $\Pr(Y_1 = j, Y_2 = k, M = 1) = \pi_{jk}\phi_{jk}$ and $\Pr(Y_1 = j, Y_2 = k, M = 0) = \pi_{jk}(1 - \phi_{jk})$. This model has $2JK - 1$ parameters, and the data have $JK + J - 1$ degrees of freedom to estimate them: JK from the fully classified data, J from the supplementary margin, less one for the constraint that the probabilities sum to one. Hence, there are $2JK - 1 - (JK + J - 1) = J(K - 1)$ parameters, too many for unique ML estimates in the unrestricted (saturated) model. We seek to reduce the number of parameters by placing hierarchical loglinear model restrictions on the cell probabilities. (Note that the loglinear models in Section 13.4 concerned the joint distribution of the Y s, whereas here we are modeling the joint distribution of both the Y s and the binary missingness indicator, M .)

All the hierarchical models that include the main effects of Y_1 , Y_2 , and M are displayed in Table 15.10. The first column describes the model using the notation introduced in Section 13.4. The next three columns give the number of parameters in the model, the number of degrees of freedom for testing the fit of the model, and the number of parameters in the model that are inestimable, because they do not appear in the likelihood. These quantities satisfy the relationship:

$$\text{df(model)} + \text{df(lack of fit)} - \text{df(inestimable)} = JK + J - 1,$$

the degrees of freedom in the data. The remaining six columns show fits to the data in Table 15.9 – the likelihood ratio (LR) chi-squared statistic for lack of fit, its associated degrees of freedom, and estimates of the cell probabilities ($\times 100$).

We note the following properties of the models in Table 15.10

1. *Inestimability*: The models $\{Y_1 Y_2 M\}$, $\{Y_1 Y_2, Y_1 M, Y_2 M\}$, $\{Y_1 M, Y_2 M\}$, $\{Y_1, Y_2 M\}$, and, if $K > J$, $\{Y_1 Y_2, Y_2 M\}$ have parameters that do not appear in their respective likelihoods. Additional information is needed to obtain unique ML estimates of the cell probabilities for these models, so they are not included in the table.

Note that two of these models, $\{Y_1 M, Y_2 M\}$ and $\{Y_1, Y_2 M\}$ have inestimable parameters, even though they have fewer parameters than degrees of freedom, $JK + J - 1$, in the data. For example, consider the model for conditional independence of Y_1 and Y_2 given M , namely, $\{Y_1 M, Y_2 M\}$. The model has $2J + 2K - 3$ parameters – one for the marginal probability of missingness, $J + K - 2$ for the conditional distribution of Y_1 and Y_2 given $M = 1$, and $J + K - 2$ for the conditional distribution of Y_1 and Y_2 given $M = 0$. The latter two distributions both have $JK - 1$ probabilities, which are subject to $(J - 1)(K - 1)$ constraints because Y_1 and Y_2 are independent, given M . The incomplete data likelihood factors into three

Table 15.10 Example 15.17, models for a two-way table with one supplemental margin

Model	Model	Degrees of freedom			Lack of fit			Example from Table 15.7		
		Inestimable	LRT	df	π_{11}	π_{12}	π_{21}	π_{22}		
(1) $\{Y_1 Y_2 M\}$	$2JK - 1$	0	$J(K-1)$	—	—	—	—	—	—	—
(2) $\{Y_1 Y_2, Y_1 M, Y_2 M\}$	$JK + J + K - 2$	0	$K-1$	—	—	—	—	—	—	—
(3) $\{Y_1 Y_2, Y_1 M\}$	$JK + J - 1$	0	0	0	44.4	8.9	17.5	29.2		
(4) $\{Y_1 Y_2, Y_2 M\}$	$JK + K - 1$	$\max(J-K, 0)$	$\max(K-J, 0)$	0	39.4	14.0	11.8	34.9		
(5) $\{Y_1 Y_2, M\}$	JK	$J-1$	0	10.75	1	44.4	8.9	17.5	29.2	
(6) $\{Y_1 M, Y_2 M\}$	$2(J+K) - 3$	$(J-1)(K-1)$	$K-1$	44.99	1	—	—	—		
(7) $\{Y_1 M, Y_2\}$	$2J + K - 2$	$(J-1)(K-1)$	0	44.99	1	34.7	18.7	30.3	16.3	
(8) $\{Y_1 Y_2, M\}$	$2K + J - 2$	$(J-1)K$	$K-1$	55.74	2	—	—	—		
(9) $\{Y_1 Y_2, M\}$	$J + K - 1$	$(J-1)K$	0	55.74	2	34.7	18.7	30.3	16.3	

components with distinct parameters, corresponding to the marginal distribution of M , the conditional distribution of Y_1 and Y_2 given $M=0$, and the conditional distribution of Y_1 given $M=1$. These three components provide estimates of $1+(J+K-2)+(J-1)=2J+K-2$ parameters; the remaining $K-1$ parameters in the model, corresponding to the distribution of Y_2 given $M=1$, are inestimable. This accounting leaves $(JK+J-1)-(2J+K-2)=(J-1)(K-1)$ degrees of freedom in the data, which correspond to lack of fit of the model with conditional independence of Y_1 and Y_2 given $M=1$.

2. *Missingness mechanism:* The models $\{Y_1 Y_2, Y_1 M\}$ and $\{Y_2, Y_1 M\}$ are MAR because missingness depends only on Y_1 , which is fully observed. These models can be fitted using the methods of Chapter 13. The models $\{Y_1 Y_2, M\}$ and $\{Y_1, Y_2, M\}$ assume the data are MCAR, and yield the same estimates of $\{\pi_{jk}\}$ as their MAR counterparts, $\{Y_1 Y_2, Y_1 M\}$ and $\{Y_2, Y_1 M\}$ respectively.
3. *Lack of fit:* The LR test for $\{Y_1 Y_2, M\}$ is based on a test of independence of Y_1 and M , using the $Y_i \times M$ two-way margin. The LR test for $\{Y_1 M, Y_2\}$ is based on a test of independence of Y_1 and Y_2 , using the fully classified data. The LR test for $\{Y_1, Y_2, M\}$ is found by summing the LR test statistics for $\{Y_1 Y_2, M\}$ and $\{Y_1 M, Y_2\}$.
4. *Estimation:* The ML estimates of $\{\pi_{jk}\}$ for $\{Y_1 Y_2, Y_1 M\}$ or $\{Y_1 Y_2, M\}$ are $\pi_{jk} = (r_{jk} + \hat{m}_{jk})/(r + m)$ where $\hat{m}_{jk} = (r_{jk}/r_{j+})m_j$ is a filled-in count (cf. Eq. (13.5)). One can view this estimate as arising from distributing the partially classified counts $\{r_j\}$ into the table to match the *row* distributions $\{m_{jk}/m_{j+}\}$ of the fully observed data, as in Examples 13.1 and 13.2.

Only one of the five MNAR models in Table 15.10 yield unique ML estimates without additional restrictions, namely, $\{Y_1 Y_2, Y_2 M\}$, which can be estimated if $K \leq J$. The model supposes that missingness of Y_2 depends on the value of Y_2 but not on the value of Y_1 . The ML estimates of $\{\pi_{jk}\}$ for this model also have the form $\hat{\pi}_{jk} = (r_{jk} + \hat{m}_{jk}^*)/(r + m)$, but now the filled-in values \hat{m}_{jk}^* are such that $\hat{m}_{jk}^*/\hat{m}_{+k}^* = r_{jk}/r_{+k}$, that is, they match the *column* distributions of the fully classified data. These constraints, together with the constraints $\sum_k \hat{m}_{jk}^* = m_j$ for all j , yield $JK - K + J$ linear equations for the JK unknowns \hat{m}_{jk}^* . When $K > J$, there are fewer equations than unknown parameters, and hence, *a priori* constraints are required to define unique ML estimates $\{\hat{m}_{jk}^*\}$ (and hence $\hat{\pi}_{jk}$). When $K < J$, there are more equations than parameters, and the ML estimates \hat{m}_{jk}^* cannot satisfy the constraints exactly; the EM algorithm can be used to calculate $\{\hat{m}_{jk}^*\}$ in such cases (see, for example, Baker and Laird (1988)). When $K=J$, the JK linear equations can be solved directly, yielding ML estimates without resorting to EM iterations. In

particular, for $J=K=2$, we obtain the following equations for \hat{m}_{11}^* , \hat{m}_{12}^* , \hat{m}_{21}^* and \hat{m}_{22}^* :

$$\begin{aligned}\hat{m}_{21}^* &= \hat{m}_{11}^* r_{21}/r_{11}; & \hat{m}_{22}^* &= \hat{m}_{12}^* r_{22}/r_{12}; & \hat{m}_{11}^* + \hat{m}_{12}^* &= m_1; \\ \hat{m}_{21}^* + \hat{m}_{22}^* &= m_2\end{aligned}$$

Solving them yields $\hat{m}_{11}^* = (m_2 - m_1 r_{22}/r_{12})(r_{21}/r_{11} - r_{22}/r_{12})^{-1}$, and so on. For the data in Table 15.7 we obtain

$$\hat{m}_{11}^* = 200/11, \quad \hat{m}_{12}^* = 240/11, \quad \hat{m}_{21}^* = 60/11, \quad \hat{m}_{22}^* = 600/11,$$

which yield the estimates of $\{\pi_{jk}\}$ in row (4) of Table 15.7.

The estimates obtained from solving these linear equations can be negative, and hence not ML. Baker and Laird (1988) show that to obtain nonnegative estimates $\{\hat{m}_{jk}^*\}$, the marginal column odds $\{m_j/m_l\}$ must lie between the smallest and largest values of the column odds $\{r_{jk}/r_{lk}\}$, $k = 1, \dots, K$. In our example, $m_1/m_2 = 40/60$ lies between $r_{11}/r_{21} = 100/30$ and $r_{12}/r_{22} = 20/50$, so this condition is satisfied. If this condition is not satisfied, then the estimates need to be modified to ensure that $\hat{m}_{jk}^* \geq 0$ for all j, k . Details are given in Baker and Laird (1988).

5. *Choice between models:* It is important to note that in our example both the models $\{Y_1, Y_2, Y_1 M\}$ and $\{Y_1, Y_2, Y_2 M\}$ yield perfect fits to the data with no degrees of freedom for testing fit. Thus, it is not possible to choose between the estimates of $\{\pi_{jk}\}$ they supply, except by *a priori* reasoning about which missingness mechanism is more plausible for the data set at hand.

The ideas of this example are generalized to a two-way table with two supplementary margins in Little (1985b). In that case, indicators M_1 and M_2 are introduced for missingness in Y_1 and Y_2 , and models for the four-way table of Y_1, Y_2, M_1 , and M_2 are considered. Similar MNAR models for higher-order tables are developed in an analogous manner.

Example 15.18 Predicting Results of the Slovenian Plebiscite with Polling Data. Prior to the Slovenian Plebiscite in 1991, in which 88.5% of eligible Slovilians voted to create an independent state, the Slovenian Public Opinion Survey (SPOS) collected information on the likely outcome of that vote, because the denominator, which was the number of eligible voters, was known. Because the SPOS suffered from nonresponse, and we know the result of the plebiscite, it serves as an interesting example for assessing the performance of MAR and MNAR missingness models.

Table 15.11, taken from Rubin et al. (1996), summarizes results from the survey for three categorical variables: “Attendance” concerns whether the

Table 15.11 Example 15.18, results of Slovenian Public Opinion Survey

Secession	Attendance	Independence		
		Yes	No	Don't know
Yes	Yes	1191	8	21
	No	8	0	4
	Don't know	107	3	9
No	Yes	158	68	29
	No	7	14	3
	Don't know	18	43	31
Don't know	Yes	90	2	109
	No	1	2	25
	Don't know	19	8	96

Reproduced with permission of Taylor and Francis.

respondents said they would participate in the plebiscite, “Independence” concerns whether they would vote for independence, and “Secession” asks the respondent’s opinion on a related issue. All three questions had “Do not Know” responses that could plausibly be treated as missing data according to the definition in Section 1.2, because in this situation they do mask real responses. Recall that in the plebiscite, all eligible voters are known, and their number is used as the denominator for the percentage voting for independence; the numerator is the number of people actively voting for it – a “non-vote” from an eligible voter counts the same as a “No” vote against independence. The critical questions in Table 15.11 are thus the ones on independence and attendance, because we wish to estimate the percentage of the eligible votes who will attend the plebiscite and vote “Yes” to the independence question. The data about succession provides potentially useful covariate information for the other two questions.

The results of various approaches to address nonresponse in the SPOS are displayed in Table 15.12. The conservative approach assumed every “Do not Know” reply was really a negative reply. The complete-cases approach used only those respondents who answered all three questions, and the available-cases approach used those who answered the independence and attendance questions. The IML estimates are based on the EM algorithm applied to a saturated ignorable multinomial model for the $2 \times 2 \times 2$ data of Table 15.11, as in Section 13.3. A DA algorithm for the same model and data with a Jeffreys’ prior distribution, as in Example 6.18, yielded a posterior median that was the same as the ML estimate to within a tenth of a percent.

Table 15.12 Example 15.18, Slovenian Public Opinion Poll Survey: comparison of estimates on independence question for different methods of handling missing data

Estimation method	Yes	No	No via nonattendance
Conservative	0.694	0.306	0.192
Complete cases	0.928	0.072	0.020
Available cases	0.929	0.071	0.021
Ignorable, ML, or Bayes	0.883	0.117	0.043
MNAR	0.782	0.218	0.122
Plebiscite = truth	0.885	0.115	0.065

Reproduced with permission of Taylor and Francis.

The MNAR model in Table 15.12 assumed that nonresponse (missingness) on a question was a function of the answer to that question. More specifically, including the missing data indicators for the data of Table 15.11 leads to a 2^6 table of counts, where we saturate the model for the $2 \times 2 \times 2$ data and the $2 \times 2 \times 2$ missing-data indicators, but allow only the three interaction parameters between data and missing-data indicators corresponding to the question and its missing data-indicator.

The last row of Table 15.12 shows the results of the plebiscite that the earlier opinion poll was attempting to predict. The only estimates that are close to the actual outcome are ones based on the ignorable model, despite the fact that the MNAR model might be regarded as reasonably sensible.

In our limited experience, this is not an uncommon result. In carefully conducted surveys with good information available on nonrespondents, ignorable missing-data models are often seen to outperform MNAR models. This is not to argue that the missingness mechanisms operating in these surveys are really MAR, but rather that the formulation of MNAR models that are superior to ignorable models is very context-specific and appears not to be easy.

15.4.3 Sensitivity Analyses for Chained-Equation Multiple Imputations

Published sensitivity analyses based on MNAR models have been largely limited to the relatively simple problem where missing values are confined to a single variable. Multiple Imputation via chained equations, as described in Section 10.2.4, is a flexible method for handling a general multivariate pattern of missing data, with various variable types, assuming MAR. A relatively simple sensitivity analysis to assess deviations from MAR is to add fixed offsets to the chained equation imputations, as illustrated in the following example.

Example 15.19 *Sensitivity Analysis for Income Nonresponse in a Rotating Panel Survey.* Giusti and Little (2011) consider the treatment of missing income

data in the labor force survey of the Municipality of Florence in Italy. A random sample of individuals is drawn from the municipal register of Florence, stratified by sex, age-class, and zone of residence. The survey has a rotating panel design, where each subject enters the sample for two consecutive quarters, exits for two, and then re-enters again for two quarters, with a 50% overlap after 3 and 12 months, and a 25% overlap after 9 and 15 months. To determine this timing, each subject is randomly assigned into one of eight “panel groups.”

We focus on missing values for the questions about occupational status and earned income for employed people. Income recipiency and amount are missing for waves where individuals are not interviewed, and income amount is missing for waves where individuals are interviewed but refuse to answer the income amount question. The result is a multivariate missing-data problem with two missingness mechanisms, one by design and one by refusal, and varying sets of covariates for imputation depending on the wave of the survey. Table 15.13 summarizes the status, observed or missing, for the occupational status (Z) and the monthly income (Y) in each of the quarters and for each panel group, and Table 15.14 shows numbers and percentages of missing values of Y .

The missing data due to the rotating panel design are MAR by design, but refusal to answer questions about income amounts is often thought to be MNAR, with nonresponse considered more likely among individuals with high and low incomes than among individuals with middle incomes.

Initially, Giusti and Little (2011) multiply-imputed missing quarterly income values and missing values on occupational status and covariates using MAR chained equation methods discussed in Section 10.2.4, which allow conditioning on available covariate information, including available income data from other quarters.

A total of 25 datasets were created using the software package IVEware (Raghunathan et al. 2001). Variables in the imputation model included occupational status and log(income) in the different waves, sex, age-class, number of household members, zone of residence in the Municipality of Florence, educational level, and civil status. They also conditioned on some characteristics available for the quarters when the subject was interviewed and employed, that is the type of job (employee or self-employed), the number of household members receiving income, and the involvement in a second job. These also needed to be imputed when not available because of the rotating scheme.

To describe modifications of this MAR analysis to examine sensitivity to MNAR missingness mechanisms, let $z_{hij} = 0, 1$, $h = 1, \dots, H$, $i = 1, \dots, n_h$, $j = 1, \dots, J$ be the indicator of the occupational status for subject i in stratum h and wave j , and let y_{hij} be the corresponding (monthly net income) from a job in Euros. If a subject is not employed ($z_{hij} = 0$), then the income is zero ($y_{hij} = 0$). Define the missingness indicator m_{hij} such that $m_{hij} = 0$ if occupational status and income are observed; $m_{hij} = 1$ if occupational status and income are both missing, as when the subject belongs in a panel group that is not interviewed

Table 15.13 Example 15.19, missingness status of Z (occupational status) and Y (monthly income)

Panel Group	April 2002		July 2002		October 2002		January 2003	
	Z	Y	Z	Y	Z	Y	Z	Y
Group 1	Obs	Obs/Mis	Mis	Mis	Mis	Mis	Mis	Mis
Group 2	Mis	Mis	Obs	Obs/Mis	Mis	Mis	Mis	Mis
Group 3	Mis	Mis	Mis	Mis	Obs	Obs/Mis	Mis	Mis
Group 4	Mis	Mis	Mis	Mis	Mis	Mis	Obs	Obs/Mis
Group 5	Obs	Obs/Mis	Mis	Mis	Mis	Mis	Obs	Obs/Mis
Group 6	Obs	Obs/Mis	Obs	Obs/Mis	Mis	Mis	Mis	Mis
Group 7	Mis	Mis	Obs	Obs/Mis	Obs	Obs/Mis	Mis	Mis
Group 8	Mis	Mis	Mis	Mis	Obs	Obs/Mis	Obs	Obs/Mis

Obs, observed; Mis, missing; Obs/Mis, some of each, by Quarter and Panel Group.

Source: Giusti and Little (2011). Reproduced with permission of Journal of Official Statistics.

Table 15.14 Example 15.19, number employed (N) and percentage of employed missing monthly income Y (% missing)

Panel Group	April 2002		July 2002		October 2002		January 2003	
	N	% missing	N	% missing	N	% missing	N	% missing
Group 1	286	31.47	0	0	0	0	0	0
Group 2	0	0	195	37.95	0	0	0	0
Group 3	0	0	0	0	174	36.21	0	0
Group 4	0	0	0	0	0	0	272	39.34
Group 5	118	31.36	0	0	0	0	119	26.05
Group 6	244	24.59	245	31.43	0	0	0	0
Group 7	0	0	239	38.49	239	36.82	0	0
Group 8	0	0	0	0	263	36.50	264	31.44
Total	648	28.86	679	35.79	676	36.54	655	33.74

Zeros in the table derive from the survey rotation scheme.

Source: Giusti and Little (2011). Reproduced with Permission of Journal of Official Statistics.

in wave j ; and $m_{hij} = 2$ if occupational status is observed but income is missing, as when an individual is interviewed but refuses to answer the income question. The MNAR mechanism is modeled via the joint distribution of y_{hij} , z_{hij} , and m_{hij} given the observed variables, which we write generically as $C_{(0), hij}$. We first factor this distribution as follows:

$$f[y_{hij}, z_{hij}, m_{hij} | C_{(0), hij}] = f[y_{hij}, z_{hij} | m_{hij}, C_{(0), hij}] \times f[m_{hij} | C_{(0), hij}],$$

which is a pattern-mixture factorization of the joint distribution (the notation here suppresses the dependence of the distributions on parameters). We assume

$$f[y_{hij}, z_{hij} | m_{hij} = 1, C_{(0), hij}] = f[y_{hij}, z_{hij} | m_{hij} \neq 1, C_{(0), hij}],$$

which expresses the fact that the distribution of y_{hij}, z_{hij} is the same for individuals who are or are not interviewed because of the rotation group design. Further, for the missing income values due to refusal, we assume that

$$f[y_{hij} | z_{hij} = 1, m_{hij} = 2, C_{(0), hij}] \neq f[y_{hij} | z_{hij} = 1, m_{hij} = 0, C_{(0), hij}],$$

which is a MNAR model because the distribution of y_{hij} given z_{hij} and $C_{(0), hij}$ is allowed to differ for nonrespondents and respondents. Note that this distribution conditions on z_{hij} because this is observed for units with $m_{hij} = 0$ or 2. Specifically, we model the difference by assuming

$$E[\log(y_{hij}) | z_{hij} = 1, m_{hij} = 2, C_{(0), hij}] = E[\log(y_{hij}) | z_{hij} = 1, m_{hij} = 0, C_{(0), hij}] + k\sigma_{hj},$$

where σ_{hj} is the residual standard deviation of the distribution of $\log(y_{hij})$ for respondents given $z_{hij} = 1$ and $C_{(0), hij}$, and k is a positive predetermined multiplier. The effect is to increase the mean of the distribution for nonrespondents relative to that for respondents by a value $k\sigma_{hj}$, which depends on the choice of k and the predictive power of $C_{(0), hij}$, as reflected in the residual standard deviation σ_{hj} . Note that the shift in the distribution for nonrespondents is applied after fitting the MAR model and is not part of the imputation algorithm. This is because we do not want the increment to be amplified by the iterations of the imputation scheme, a point discussed in Van Buuren et al. (1999). This model is implemented as follows:

- (A) The MAR multiple imputations were created as before;
- (B) A value of k (0.8, 1.2, or 1.6) is chosen to reflect small, medium, and large deviations from MAR. The offsets are then applied to the imputations for refusals;
- (C) For each of the m sets of multiple imputations, the imputations for the refusals are treated as known, and the sequential multiple imputation method is applied to re-impute the missing values of Y and Z for months

not in the rotation group. This approach allows these imputations to condition on the offsetted values of the nonrespondents, reflecting the fact that individuals not in the rotation group may also fail to respond if interviewed.

Giusti and Little (2011) label this imputation model MNAR₁. They also present results under an alternative assumption (denoted MNAR₂), where missing values for units with at least one income value reported are regarded as MAR, and the offsetting restricted to units with no observed income values. The MNAR₂ mechanism is clearly closer to MAR than model MNAR₁, and MNAR₁ and MNAR₂ can be thought of as bounding a range of plausible combinations of these models, for any choice of k .

To evaluate the impact of the MNAR increments on the income distributions referring to the four quarters, Figure 15.3 plots empirical densities from one set of imputed income values for the first quarter, compared with the density of the corresponding observed values, under the MNAR₁ model. The plots show the impact of the proposed MNAR imputation models on the income distribution in April. As expected, larger k values cause a more pronounced shift for the corresponding density. Plots under the MNAR₂ model show a reduced shift for the distributions relative to those under MNAR₁.

For the value $k = 0.8$, the percentage increase of the quarterly income estimates relative to the MAR model is around 10% for the MNAR₁ model and 7% for the MNAR₂ model. For $k = 1.2$ and $k = 1.6$, greater percentage increases are generated. Comparisons of results with external estimates from a national survey conducted by the Italian National Institute of Statistics (ISTAT) suggest that the value $k = 1.6$ can be considered as a plausible maximum for our proposed MNAR models. Broadly speaking, we can say that the impact of MNAR deviations from the MAR estimates are moderate, especially under the MNAR₂ model.

15.4.4 Sensitivity Analyses in Pharmaceutical Applications

In the world of pharmaceutical development, randomized experiments with humans play a major role for the approval of products for commercial sale in many parts of the world. For example, in the United States, the US Food and Drug Administration (US FDA) relies on such experiments before approving most products. Accompanying such experiments typically are missing data, sometimes in primary outcomes and commonly in secondary outcomes used to create a more complete view of the overall medical benefit of the product being considered. In this environment, a simple way to deal with missing data is to do a “worst-case” analysis, where for instance, all people with missing data in the active treatment arm are assumed to be failures (e.g., have died), whereas all people with missing data in the control are assumed to be successes

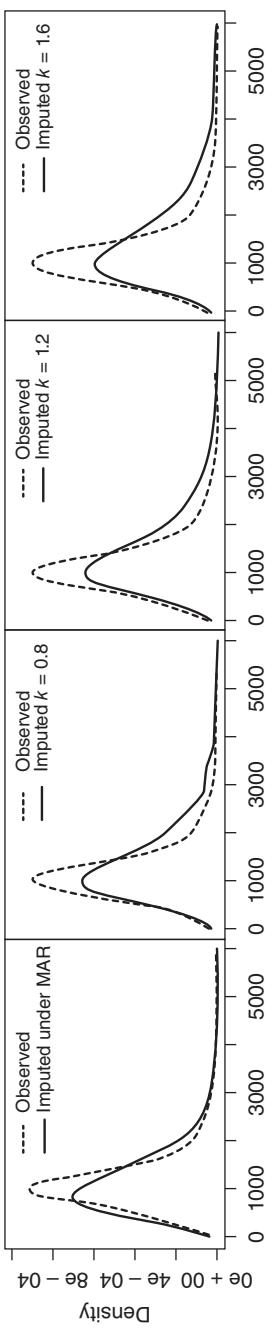


Figure 15.3 Example 15.19, empirical densities of the observed income values in the first quarter (dotted lines) and of imputed income values (solid lines) under the MNAR₁ model. Source: Giusti and Little (2011). Reproduced with Permission of Journal of Official Statistics.

(e.g., have survived). If the product still appears successful according to the protocol-defined assessments, the occurrence of the missingness is considered unimportant for the product's approval. But in practice, such an extreme assumption leads to more ambiguous conclusions, and is not considered scientifically reasonable, especially when there exists historical evidence contradicting such an extreme analysis.

A National Research Council (NRC) study (Little et al. 2012; National Research Council 2010) provided recommendations for limiting missing data in clinical trial design and conduct, and principles for analysis, including the need for sensitivity analyses to assess robustness of findings to alternative assumptions about the missing data. Our final two examples describe a sensitivity analysis in the context of survival analysis, and graphical methods for displaying sensitivity.

Example 15.20 *A Sensitivity Analysis to Assess the Potential Impact of Differential Nonignorable Censoring in Survival Analysis.* This example summarizes analyses to address concerns about missing data raised by a FDA advisory committee, in their review of results from the ATLAS ACS 2 TIMI 51 study (Mega et al. 2012), a large randomized, double-blind, placebo-controlled clinical trial that assessed rivaroxaban for its ability to reduce the risk of cardiovascular death, myocardial infarction or stroke in patients with acute coronary syndrome. For more details, see Little et al. (2016b).

A total of 15 526 patients were randomized in a 1 : 1 : 1 ratio into three treatment groups: rivaroxaban 2.5 mg twice a day, rivaroxaban 5 mg twice a day, and placebo. The primary efficacy outcome was a composite of cardiovascular death, myocardial infarction or stroke. Two forms of analysis were prespecified in the protocol; we present results from both here because they differ greatly in their amounts of missing data. The primary efficacy analysis included outcomes of all randomized participants up to the earlier date of the day before global treatment end date, 30 days after the last study treatment, or 30 days after randomization for those who had not received any study medication. This analysis was called "modified intent-to-treat" (mITT), because of the restriction to events within 30 days of last treatment or randomization. Strict intent-to-treat (ITT) analyses were also carried out, which included all events occurring up until the global treatment end date. For the primary mITT analysis, the study showed a reduction in hazard for the combined rivaroxaban groups relative to control (hazard ratio (HR) = 0.84, 95% confidence interval (CI) = (0.74, 0.96)). The ITT results were slightly more favorable for rivaroxaban: HR = 0.82, 95% CI = (0.73, 0.93). The key safety endpoint was non-CABG TIMI major bleeding, which was increased with rivaroxaban treatment (HR = 3.96, 95% CI = (2.46, 6.38)).

Despite these positive results, the FDA Advisory Committee voted against approval, and the issue of missing data was a primary concern. The primary endpoint was missing for 799 (5.1%) of the 15 526 participants for the primary mITT outcome, and missing for 1509 (9.7%) of participants for the ITT outcome. Reasons for being missing were classified as “adverse event,” “consent withdrawn,” “lost to follow up,” or “other,” which included subjects who were randomized but did not meet inclusion and exclusion criteria. The main concern raised by missing data is that individuals who discontinue often differ systematically from units who complete; the concern is particularly important if these differences differ by treatment group. As discussed in Section 6.4, censoring by treatment discontinuation is CAR if subjects with missing follow-up have the same hazards as those with complete follow-up, after adjusting for observed data up to time of loss of follow-up. Otherwise, we say the censoring is coarsened not at random (CNAR). One hypothetical scenario of CNAR occurs if participants with missing follow-up have a high rate of bleeding before dropping out, and a high rate of bleeding leads to a higher chance of subsequent cardiovascular events. CNAR censoring is *differential* if it leads to bias in the comparison of rivaroxaban and placebo groups; that is, if the differences in the outcomes due to CNAR censoring in the treatment groups do not “cancel out.”

The amount of missing data appears much higher when measured by the *fraction of units* with missing outcomes (5.1% for mITT, 9.7% for ITT) than when measured by the *fraction of person-years* with missing outcomes (0.3%, 6.9%). The differences between mITT and ITT illustrate an observation in the NRC report that the amount of missing data can vary substantially depending on the choice of primary estimand, and limiting missing data should be a consideration in the choice of estimand. Little et al. (2016b) propose the *fraction of missing information* for the particular estimand, as discussed in Section 10.2, as a more principled measure, a measure that was closer to the fraction of person-years than the fraction of units in this example.

In response to the concerns, the following steps were taken to assess the potential impact of the missing data on the results of the trial. First, key baseline characteristics and clinical events preceding withdrawal of consent or discontinuation from the study were assessed for patients who dropped out, completed the study, or died during the study. Second, the robustness of the ATLAS ACS 2 TIMI 51 study results was assessed using a pattern-mixture model analysis. As in standard implementations of survival analysis, the primary analyses here assume the censoring mechanism is CAR. To assess deviations from CAR censoring, survival times for participants in the rivaroxaban group who withdrew from the study were imputed based on hazards that were allowed to deviate by prespecified multiples from those of participants who did not withdraw. Uncertainty inherent in imputation was propagated by multiple imputation. Third, the sponsor undertook an intense global effort to gather information from the participants whose data on vital status were

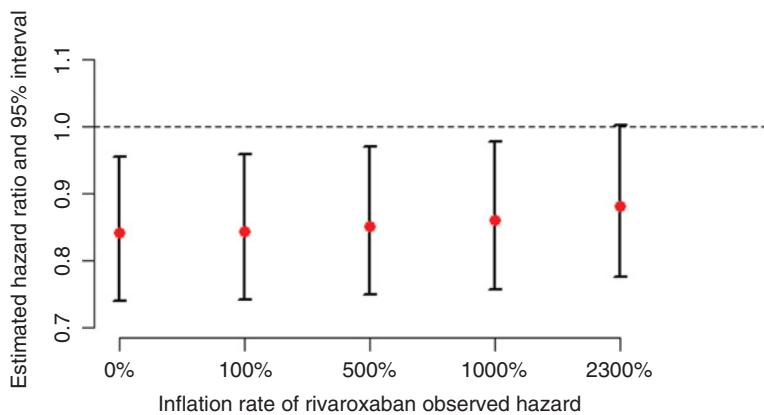
missing. A combination of site-directed activities – contacting participants or checking paper and electronic medical records – and national database queries were performed to determine vital status on as many participants as possible. The originally reported findings on mortality were compared with results that included the vital status information obtained in the follow-up study. We focus here on the sensitivity analysis.

As described in Little et al. (2016b), the sponsors fit Weibull survival models to impute outcomes of participants who (i) withdrew consent, (ii) experienced adverse events, and (iii) were lost-to-follow-up and discontinued prematurely for other reasons, with predictor variables for the primary efficacy endpoint consisting of treatment group (rivaroxaban doses pooled and placebo), and a set of baseline covariates, including demographic variables and variables reflecting previous heart disease. The model also included time-varying bleeding indicators.

The hazard at time of dropout was then inflated by a prespecified factor, for participants in the rivaroxaban groups who withdrew consent and prematurely discontinued from the study without providing a primary outcome. The corresponding hazards for participants who dropped out in the control group were not inflated, that is, for these participants, dropout was treated as CAR. The resulting hazards were used to impute events to the end of the study 1000 times, assuming a Weibull distribution. A Cox proportional-hazards model was then fitted to each of the completed datasets, and inferences for parameters combined using standard multiple-imputation combining rules described in Section 10.2. The inflation factor was increased until the upper limit of the 95% CI for the hazard ratio for rivaroxaban relative to control reached 1.0 – the “tipping point.” Technical details on the sensitivity analysis are provided in Appendix 1 of Little et al. (2016b).

The results of this sensitivity analysis are displayed in Figure 15.4 for the mITT population, where the tipping point was 2300%, and in Figure 15.5 for the ITT population, where the tipping point was 160%. The much lower tipping point for the ITT population reflects the much greater extent of imputation in the ITT analysis, because events for participants who discontinued early in the study were imputed for the entire period up to global treatment end date, rather than only for the month after discontinuation, as was done for the mITT analysis. The sponsors argued, based on this analysis, that findings of the ATLAS ACS 2 TIMI 51 study were robust to missing data; this robustness is reinforced by the follow-up study, because inclusion of data from this study had little impact on the conclusions.

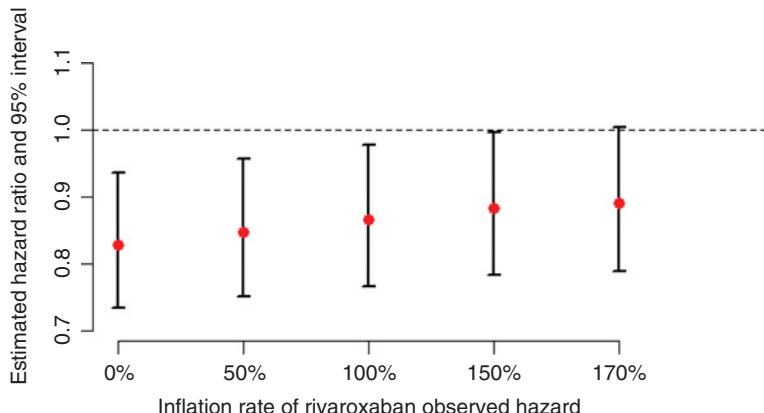
Example 15.21 Enhanced Tipping-Point Displays. An extension of the tipping-point approach in Example 15.20, called “Enhanced Tipping-Point Displays,” was proposed in Liublinska and Rubin (2012). To create these displays, models were built to multiply impute the missing outcomes. Starting with a MAR model, 100 values were imputed, thereby creating 100 completed



Mean number of imputed events:

Rivaroxaban	3	4	10	17	34
Placebo	2	2	2	2	2

Figure 15.4 Example 15.20, hazard ratio and 95% intervals for combined rivaroxaban vs. placebo, mITT analysis of primary outcome. Sensitivity analysis, inflating the individually estimated hazard in the rivaroxaban groups by known factors. Tipping point for no effect = 2300%. Source: Little et al. 2016b. Reproduced with permission of SAGE.



Mean number of imputed events:

Rivaroxaban	41	58	73	88	94
Placebo	21	21	21	21	21

Figure 15.5 Example 15.20, hazard ratio and 95% intervals for combined rivaroxaban vs. placebo, ITT analysis of primary outcome. Sensitivity analysis, inflating the individually estimated hazard in the rivaroxaban groups by known factors. Tipping point for no effect = 160%. Source: Little et al. 2016b. Reproduced with permission of SAGE.

data sets and 100 possible answers based on the number of successes/failures imputed in each arm. Focusing on the most extreme of the 100 imputations in each arm creates upper and lower bounds in each arm, and thereby on each axis in a tipping point display, where the resulting cell in the enhanced tipping point display can present any summary statistics, such as a p -value or a point estimate.

Figure 15.6 presents an example from Liublinska and Rubin (2012), where each different color represents a different MI model; the tick marks on the axes represent historical values from prior studies of related treatments. Roughly speaking, each displayed box represents a 99% interval for the possible conclusions under the associated model for the missing data, and the collection of boxes represents the sensitivity of conclusions to the assumed models for the missingness.

The main advantages of such displays are that each outcome variable will have its own display, and each display reveals the results of different modeling assumptions. Using modern computing environments, for each outcome variable, many models can be considered, and details of models that yield untoward results can be investigated, whereas models that yield benign conclusions can be ignored. Current “point and click” software and electronic reports can make such displays of dozens of models applied to dozens of outcomes easy to report, present and evaluate.

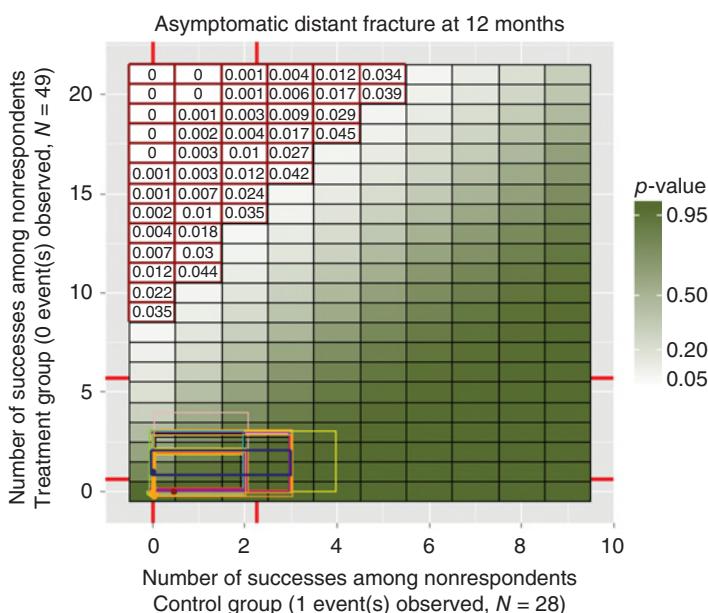


Figure 15.6 Example 15.21, an enhanced tipping-point display from Liublinska and Rubin (2012). Reproduced with permissions of American Statistical Association.

For a specific example, Liublinska and Rubin (2012) considered one MAR model and 16 non-MAR models, for six secondary outcomes in a study of a new device for the surgical treatment of osteoporosis relative to the traditional device. The report submitted to the FDA included such displays for 12 secondary outcomes. Having many such displays suggests that attention will be focused only on those displays that have models and outcomes needing attention from those approving the drugs or devices. Because of modern technology, literally hundreds of such displays can be presented and considered today, whereas doing so using printed materials would be hopelessly tedious.

The extension of these ideas to more complex situations, for example with continuous outcomes, is a topic of current development.

Problems

- 15.1** Carry out the integrations needed to derive the E step in Example 15.3.
- 15.2** Derive the expressions for the E step in Example 15.4. Also display the M step for this example explicitly.
- 15.3** Derive the E and M steps of the EM algorithm for the model in Example 15.7.
- 15.4** Review the two-step fitting method for the model of Example 15.7 of Heckman (1976). Contrast the assumptions made by that method and by the ML fitting procedure in Problem 15.3 (see e.g., Little 1985a).
- 15.5** Suppose that for the model of Example 15.7, a random subsample of nonrespondents to Y_1 is followed up and values of Y obtained. Write the likelihood for the resulting data and describe the E and M steps of the EM algorithm.
- 15.6** Derive the expressions for the posterior mean and variance of the finite population mean estimand \bar{y} in Example 15.10. What is the posterior mean and variance of variable 32D when $\psi_1 = \psi_2 = 0.5$?
- 15.7** In Example 15.13, it is shown that, for the pattern-mixture model (15.15) with MAR restrictions (15.18), the ML estimate of μ_2 is the same as for the ignorable selection model in Section 7.2.1. Show explicitly that this statement also applies for the mean of Y_1 and the covariance matrix of (Y_1, Y_2) .
- 15.8** Fill in the details leading to the ML estimates (15.22)–(15.24) for the pattern-mixture model (15.15) under restrictions (15.20).

- 15.9** Fill in the details leading to the ML estimates (15.27)–(15.29) for the pattern-mixture model (15.15) under restriction (15.26).
- 15.10** Show that if $\lambda = -\beta_{12 \cdot 2}^{(0)}$, substituting the ML estimate of $\beta_{12 \cdot 2}^{(0)}$ in Eqs. (15.27)–(15.29) yields complete-case estimates. That is, if $\lambda = -\beta_{12 \cdot 2}^{(0)}$ is thought to be more plausible than $\lambda = 0$, then the complete case estimate of μ_2 is better than the ML estimate assuming ignorable nonresponse.
- 15.11** For the pattern-mixture model (15.15) where missingness of Y_2 depends on $Y_1 + \lambda Y_2$, show that the ML estimate of $c_1\mu_1 + c_2\mu_2$ is $c_1\bar{y}_1 + c_2\bar{y}_2 + (c_1 + c_2 b_{21 \cdot 1}^{(\lambda)})(\hat{\mu}_1 - \bar{y}_1)$. Hence, show that the ML estimate of $\mu_1 - \mu_2$ is the complete case estimate, $\bar{y}_1 - \bar{y}_2$, when $\lambda = (\sigma_{11}^{(0)} - \sigma_{12}^{(0)})/(\sigma_{22}^{(0)} - \sigma_{12}^{(0)})$; the IML estimate when $\lambda = 0$; and the available case estimate $\hat{\mu}_1 - \bar{y}_2$ when $\lambda = -\beta_{12 \cdot 2}^{(0)}$. Deduce situations where the available case estimate is the preferred estimate (see Little 1994).
- 15.12** For suitable parameterizations of the models, write down factored likelihoods for the models $\{Y_1 Y_2, Y_1 M\}$, $\{Y_1 Y_2, Y_2 M\}$, $\{Y_1 M, Y_1 M\}$, $\{Y_1, Y_2 M\}$ in Example 15.17. State for each model which parameters (if any) are inestimable, in the sense that they do not enter the likelihood.
- 15.13** Verify the five sets of estimated cell probabilities in Table 15.10.
- 15.14** Redo Table 15.10 for the data in Table 15.9 with m_1 and m_2 multiplied by a factor of 10.
- 15.15** Reproduce the EM estimates for the ignorable model in Table 15.12. Use the bootstrap to estimate the standard error of the estimated proportion voting yes, and compare it with the large sample standard error.
- 15.16** Reproduce the Bayes' estimates for the ignorable model in Table 15.12 by data augmentation, and provide a histogram of draws. Use 10 draws of the missing data under DA to create multiple imputations of the missing data, and apply the general combining rules to draw an inference for the percent voting yes. Compare this answer with the corresponding answers in Problem 15.15.
- 15.17** Repeat the calculations in Problem 15.16 for the MNAR model in Table 15.12.

References

- Aboyomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multiple imputations. *Appl. Stat.* 57 (3): 273–291.
- Afifi, A.A. and Elashoff, R.M. (1966). Missing observations in multivariate statistics 1: review of the literature. *J. Am. Stat. Assoc.* 61: 595–604.
- Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55: 117–128.
- Aitkin, M. and Rubin, D.B. (1985). Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. B* 47: 67–75.
- Aitkin, M. and Wilson, G.T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics* 22: 325–331.
- Albert, P.S., Follman, D.A., Wang, S.A., and Suh, E.B. (2002). A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics* 58 (3): 631–664.
- Allan, F.G. and Wishart, J. (1930). A method of estimating the yield of a missing plot in field experiments. *J. Agric. Sci.* 20: 399–406.
- Amemiya, T. (1984). Tobit models: a survey. *J. Econom.* 24: 3–61.
- Anderson, R.L. (1946). Missing plot techniques. *Biometrics* 2: 41–47.
- Anderson, T.W. (1957). Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *J. Am. Stat. Assoc.* 52: 200–203.
- Anderson, T.W. (1965). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.
- Andrews, D.F., Bickel, P.J., Hampel, F.R. et al. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, NJ: Princeton University Press.
- Andridge, R. and Little, R.J. (2009). Extensions of proxy pattern-mixture analysis for survey nonresponse. In: *Proceedings of the Survey Research Methods Section, 2009*, 2468–2482. American Statistical Association.
- Andridge, R.H. and Little, R.J. (2010). A review of hot deck imputation for survey nonresponse. *Int. Stat. Rev.* 78 (1): 40–64.

- Andridge, R.H. and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *J. Off. Stat.* 27 (2): 153–180.
- Angrist, I.D., Imbens, G.W., and Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91: 444–472 (with discussion).
- Azen, S. and Van Guilder, M. (1981). Conclusions regarding algorithms for handling incomplete data. In: *Proceedings of the Statistical Computing Section, 1981*, 53–56. American Statistical Association.
- Bailar, B.A. and Bailar, J.C. (1983). Comparison of the biases of the “hot deck” imputation procedure with an “equal weights” imputation procedure. In: *Incomplete Data in Sample Surveys: Symposium on Incomplete Data, Proceedings*, vol. 3 (ed. W.G. Madow and I. Olkin). New York: Academic Press.
- Bailar, B.A., Bailey, L., and Corby, C. (1978). A comparison of some adjustment and weighting procedures for survey data. In: *Proceedings of the Survey Research Methods Section, 1978*, 175–200. American Statistical Association.
- Baker, S.G. and Laird, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Am. Stat. Assoc.* 83: 62–69.
- Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61: 962–972.
- Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press.
- Barnard, J. and Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* 86: 949–955.
- Barnard, J., Du, I., Hill, I., and Rubin, D.B. (1998). A broader template for analyzing broken randomized experiments. *Sociol. Methods Res.* 27: 285–318.
- Bartlett, M.S. (1937). Some examples of statistical methods of research in agriculture and applied botany. *J. R. Stat. Soc. B* 4: 137–170.
- Baum, L.E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41: 164–171.
- Beale, E.M.L. and Little, R.J. (1975). Missing values in multivariate analysis. *J. R. Stat. Soc. B* 37: 129–145.
- Beaton, A.E. (1964). The Use of Special Matrix Operations in Statistical Calculus. Educational Testing Service Research Bulletin, RB-64-51.
- Becker, M.P., Yang, I., and Lange, K. (1997). EM algorithms without missing data. *Stat. Methods Med. Res.* 6: 38–54.
- Bentler, P.M. and Tanaka, J.S. (1983). Problems with EM for ML factor analysis. *Psychometrika* 48: 247–253.
- Besag, J. (1986). On the statistical analysis of dirty pictures. *J. R. Stat. Soc. B* 48: 259–279.
- Bethlehem, J.G. (2002). Weighting adjustments for ignorable nonresponse. In: *Survey Nonresponse*, Chapter 18 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little). New York: Wiley.

- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Bondarenko, I. and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Stat. Med.* 35 (17): 3007–3020.
- Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. R. Stat. Soc. B* 26: 211–252.
- Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.
- Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.
- Box, M.J., Draper, N.R., and Hunter, W.G. (1970). Missing values in multi-response nonlinear data fitting. *Technometrics* 12: 613–620.
- Box, G.E., Hunter, J.S., and Hunter, W.G. (1985). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*. New York: Wiley.
- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 88: 9–25.
- Breslow, N.E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 82: 81–91.
- Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* 46: 143–157.
- Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.
- Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Stat. Soc. B* 22: 302–306.
- Carpenter, J.R. and Kenward, M.G. (2014). *Multiple Imputation and Its Application*. New York: Wiley.
- Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *J. Am. Stat. Assoc.* 85: 652–663.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2e. Boca Raton, FL: Chapman and Hall/CRC.
- Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In: *Incomplete Data in Sample Surveys: Symposium on Incomplete Data, Proceedings*, vol. 3 (ed. W.G. Madow and I. Olkin). New York: Academic Press.
- Chaurasia, A. and Harel, O. (2015). Partial F-tests with multiply imputed data in the linear regression framework via coefficient of determination. *Stat. Med.* 34 (3): 432–443.
- Chen, T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially classified data. *Biometrics* 30: 629–642.
- Cochran, W.G. (1977). *Sampling Techniques*, 3e. New York: Wiley.
- Cochran, W.G. and Cox, G. (1957). *Experimental Design*. London: Wiley.

- Cochran, W.G. and Rubin, D.B. (1973). Controlling bias in observational studies: a review. *Sankhya A* 35: 417–446.
- Cole, S.R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* 35 (4): 1074–1081.
- Cox, D.R. (1975). Partial likelihood. *Biometrika* 62 (2): 269–276.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. New York: Wiley.
- Crainiceanu, C.M., Ruppert, D., and Wand, M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *J. Stat. Softw.* 14 (14): 1–24.
- Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from advance data using propensity modeling; an application to income and tax statistics. *J. Bus. Econ. Stat.* 10: 117–132.
- David, M.H., Little, R.J., Samuhel, M.E., and Triest, R.K. (1983). Imputation methods based on the propensity to respond. In: *Proceedings of the Business and Economic Statistics Section, 1983*, 168–173. American Statistical Association.
- David, M.H., Little, R.J., Samuhel, M.E., and Triest, R.K. (1986). Alternative methods for CPS income imputation. *J. Am. Stat. Assoc.* 81: 29–41.
- Davies, O.L. (1960). *The Design and Analysis of Industrial Experiments*. New York: Hafner.
- Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* 56: 464–474.
- DeGroot, M.H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
- DeGroot, M.H. and Goel, K. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Stat.* 8: 264–278.
- DeGruttola, V. and Tu, X.M. (1994). Modeling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* 50: 1003–1014.
- Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.
- Dempster, A.P. and Rubin, D.B. (1983). Introduction. In: *Incomplete Data in Sample Surveys: Theory and Bibliography*, vol. 2 (ed. W.G. Madow, I. Olkin and D.B. Rubin), 3–10. New York: Academic Press.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* 39: 1–38 (with discussion).
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. *Multivariate Anal.* 5: 35–37.
- Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance component models. *J. Am. Stat. Assoc.* 76: 341–353.
- Diggle, P. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc. C* 43: 49–73.
- Dodge, Y. (1985). *Analysis of Experiments with Missing Data*. New York: Wiley.
- Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. New York: Wiley.

- Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*, Lecture Notes in Statistics. New York: Springer.
- Dunson, D.B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Stat. Assoc.* 104: 1042–1051.
- Edwards, A.W.F. (1992). *Likelihood: Expanded Edition*. Baltimore, MD: Johns Hopkins University Press.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7: 1–26.
- Efron, B. (1987). Better bootstrap confidence intervals. *J. Am. Stat. Assoc.* 82: 171–200 (with discussion).
- Efron, B. (1994). Missing data, imputation, and the bootstrap. *J. Am. Stat. Assoc.* 89: 463–478.
- Efron, B. and Hinkley, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* 65: 457–487.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: CRC Press.
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Sci.* 11: 89–121.
- Ekholm, A. and Skinner, C. (1998). The Muscatine children's obesity data reanalysed using pattern mixture models. *Appl. Stat.* 47: 251–264.
- Ernst, L.R. (1980). Variance of the estimated mean for several imputation procedures. In: *Proceedings of the Survey Research Methods Section, 1980*, 716–721. American Statistical Association.
- Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., Rubin, D., and Schafer, J. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. In: *Proceedings of 1995 Annual Research Conference*, 257–266. U.S. Bureau of the Census.
- Fay, R.E. (1986). Causal models for patterns of nonresponse. *J. Am. Stat. Assoc.* 81: 354–365.
- Fay, R.E. (1992). When are inferences from multiple imputation valid? In: *Proceedings of the Survey Research Methods Section*, 227–232. American Statistical Association.
- Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Am. Stat. Assoc.* 91: 490–498.
- Fienberg, S.E. (1980). *The Analysis of Crossclassified Data*, 2e. Cambridge, MA: MIT Press.
- Firth, D. (1991). Generalized linear models. In: *Statistical Theory and Modelling: In Honour of Sir David Cox* (ed. D.V. Hinkley, N. Reid and E.J. Snell), 55–82. New York: Chapman and Hall.
- Ford, B.N. (1983). An overview of hot deck procedures. In: *Incomplete Data in Sample Surveys: Theory and Annotated Bibliography*, vol. 2 (ed. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press.

- Frangakis, C. and Rubin, D.B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika* 86: 366–379.
- Frangakis, C. and Rubin, D.B. (2001). Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics* 57: 333–353 (with discussion and rejoinder).
- Frangakis, C.E. and Rubin, D.B. (2002). Principal stratification in causal inference. *Biometrics* 58: 21–29.
- Franks, A.M., Airoldi, E.M., and Rubin, D.B. (2016). Non-standard conditionally specified models for nonignorable missing data. arXiv:1603.06045 [stat.ME].
- Freedman, L.S., Midthune, D., Carroll, R.J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Stat. Med.* 27 (25): 5195–5216.
- Frumento, P., Mealli, F., Pacini, B., and Rubin, D.B. (2016). The fragility of standard inferential approaches in complex mixture models relative to direct likelihood approaches. *Stat. Anal. Data Min.* 9: 58–70.
- Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Am. Stat. Assoc.* 77: 270–278.
- Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.
- Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85: 398–409.
- Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. Am. Stat. Assoc.* 85: 972–985.
- Gelman, A.E. and Carlin, J.B. (2002). Poststratification and weighting adjustments. In: *Survey Nonresponse*, Chapter 19 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little). New York: Wiley.
- Gelman, A.E. and Meng, X.L. (1998). Computing normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat. Sci.* 13: 163–185.
- Gelman, A.E. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7: 457–472 (with discussion).
- Gelman, A.E., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- Gelman, A.E., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3e. London: CRC Press.
- Geyer, C.J. (1992). Practical Markov Chain Monte Carlo. *Stat. Sci.* 7: 473–503 (with discussion).
- Giusti, C. and Little, R.J. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *J. Off. Stat.* 27 (2): 211–229.
- Glynn, R.J. and Laird, N.M. (1986). Regression Estimates and Missing Data: Complete-Case Analysis. *Technical Report*. Harvard School of Public Health, Department of Biostatistics.

- Glynn, R.J., Laird, N.M., and Rubin, D.B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In: *Drawing Inferences from Self-Selected Samples* (ed. H. Wainer), 115–142. New York: Springer.
- Glynn, R.J., Laird, N.M., and Rubin, D.B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *J. Am. Stat. Assoc.* 88: 984–993.
- Goodman, L.A. (1970). The multivariate analysis of qualitative data: interaction among multiple classifications. *J. Am. Stat. Assoc.* 65: 225–256.
- Goodman, L.A. (1979). Simple models for the analysis of association in crossclassifications having ordered categories. *J. Am. Stat. Assoc.* 74: 537–552.
- Goodnight, J.H. (1979). A tutorial on the SWEEP operator. *Am. Stat.* 33: 149–158.
- Goodrich, R.L. and Caines, P.E. (1979). Linear system identification from nonstationary cross-sectional data. *IEEE Trans. Autom. Control* 24: 403–411.
- Greenlees, W.S., Reece, J.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Am. Stat. Assoc.* 77: 251–261.
- Groves, R., Dillman, D., Eltinge, J., and Little, R. (eds.) (2002). *Survey Nonresponse*. New York: Wiley.
- Guo, Y. and Little, R.J.A. (2011). Regression analysis with covariates that have heteroscedastic measurement error. *Stat. Med.* 30 (18): 2278–2294.
- Guo, Y., Little, R.J., and McConnell, D.S. (2011). On using summary statistics from an external calibration sample to correct for covariate measurement error. *Epidemiology* 23 (1): 165–174.
- Gupta, N.K. and Mehra, R.K. (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Trans. Autom. Control* 19: 774–783.
- Haberman, S.J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.
- Haitovsky, Y. (1968). Missing data in regression analysis. *J. R. Stat. Soc. B* 30: 67–81.
- Hajek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* 4: 49–57.
- Hajek, J. (1971). Comment on “An essay on the logical foundations of survey sampling, part one”. In: *The Foundations of Survey Sampling* (ed. V.P. Godambe and D.A. Sprott), 236. Holt, Rinehart, and Winston.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
- Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*, vol. 1 and 2. New York: Wiley.
- Hanson, R.H. (1978). The Current Population Survey: Design and Methodology. *Technical Paper No. 40*. U.S. Bureau of the Census.
- Harel, O. (2009). The estimation of R^2 and adjusted R^2 in incomplete data sets using multiple imputation. *J. Appl. Stat.* 36 (10): 1109–1118.

- Hartley, H.O. (1956). Programming analysis of variance for general-purpose computers. *Biometrics* 12: 110–122.
- Hartley, H.O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics* 14: 174–194.
- Hartley, H.O. and Hocking, R.R. (1971). The analysis of incomplete data. *Biometrics* 27: 783–808.
- Hartley, H.O. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54: 93–108.
- Harvey, A.C. (1981). *Time Series Models*. New York: Wiley.
- Harvey, A.C. and Phillips, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika* 66: 49–58.
- Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* 72: 320–340 (with discussion).
- Hasselblad, V., Stead, A.G., and Galke, W. (1980). Analysis of coarsely grouped data from the lognormal distribution. *J. Am. Stat. Assoc.* 75: 771–778.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109.
- Healy, M.J.R. and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Appl. Stat.* 5: 203–206.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Ann. Econ. Soc. Meas.* 5: 475–492.
- Heeringa, S.G., Little, R.J., and Raghunathan, T. (2002). Multivariate imputation of coarsened survey data on household wealth. In: *Survey Nonresponse*, Chapter 24 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little). New York: Wiley.
- Heitjan, D.F. (1994). Ignorability in general incomplete-data models. *Biometrika* 81 (4): 701–708.
- Heitjan, D.F. and Little, R.J. (1991). Multiple imputation for the fatal accident reporting system. *Appl. Stat.* 40: 13–29.
- Heitjan, D.F. and Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Am. Stat. Assoc.* 85 (410): 304–314.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31: 423–447.
- Herzog, T. and Rubin, D.B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In: *Incomplete Data in Sample Surveys: Theory and Bibliography*, vol. 2 (ed. W.G. Madow, I. Olkin and D.B. Rubin), 209–245. New York: Academic Press.
- Higgins, K.M., Davidian, M., Chew, G., and Burge, H. (1998). The effect of serial dilution error on calibration inference in immunoassay. *Biometrics* 54: 19–32.

- Hirano, K., Imbens, G., Rubin, D.B., and Zhou, X.H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1: 69–88.
- Hocking, R.R. and Oxspring, H.H. (1974). The analysis of partially categorized contingency data. *Biometrics* 30: 469–483.
- Holland, P.W. (1986). A comment on remarks by Rubin and Hartigan. In: *Drawing Inferences from Self-Selected Samples* (ed. H. Wainer), 149–151. New York: Springer.
- Holland, P.W. and Wightman, L.E. (1982). Section pre-equating: a preliminary investigation. In: *Test Equating* (ed. P.W. Holland and D.B. Rubin). New York: Academic Press.
- Holt, D. and Smith, T.M.F. (1979). Post stratification. *J. R. Stat. Soc. A* 142: 33–46.
- Horton, N.J. and Laird, N.M. (1998). Maximum likelihood analysis of generalized linear models with missing covariates. *Stat. Methods Med. Res.* 8: 37–50.
- Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *J. Am. Stat. Assoc.* 47: 663–685.
- Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 221–233. University of California Press.
- Ibrahim, J.G. (1990). Incomplete data in generalized linear models. *J. Am. Stat. Assoc.* 85: 765–769.
- Ibrahim, J.G., Lipsitz, S.R., and Chen, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. B* 61: 173–190.
- Ireland, C.T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* 55: 179–188.
- Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Stat.* 23 (3): 774–786.
- Jamshidian, M. and Jennrich, R.I. (1993). Conjugate gradient acceleration of the EM algorithm. *J. Am. Stat. Assoc.* 88: 221–228.
- Janssens, W., van der Gaag, J., Rinke de Wit, T.F., and Tanović, Z. (2014). Refusal bias in the estimation of HIV prevalence. *Demography* 51 (3): 1131–1157.
- Jarrett, R.G. (1978). The analysis of designed experiments with missing observations. *Appl. Stat.* 27: 38–46.
- Jennrich, R.I. and Schluchter, M.D. (1986). Incomplete repeated-measures models with structured covariance matrices. *Biometrics* 42: 805–820.
- Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* 22: 389–395.
- Jurek, A.M., Maldonado, G., Greenland, S., and Church, T.R. (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur. J. Epidemiol.* 21 (12): 871–876.
- Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* 82: 34–35.

- Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. In: *Proceedings of the Survey Research Methods Section 1981*, 146–151. American Statistical Association.
- Kang, J.D.Y. and Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat. Sci.* 22 (4): 523–539.
- Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.
- Kennickell, A.B. (1991). Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation. In: *Proceedings of the Section on Survey Research Methods*, 1–10. American Statistical Association.
- Kent, J.T., Tyler, D.E., and Vardi, Y. (1994). A curious likelihood identity for the multivariate t-distribution. *Commun. Stat. B – Simul. Comput.* 23: 441–453.
- Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Stat. Sci.* 13: 236–247.
- Kim, J.O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociol. Methods Res.* 6: 215–240.
- Kish, L. (1992). Weighting for unequal P_i . *J. Off. Stat.* 8: 183–200.
- Kleinbaum, D.G., Morgenstern, H., and Kupper, L.L. (1981). Selection bias in epidemiological studies. *Am. J. Epidemiol.* 113: 452–463.
- Kong, A., Liu, J.S., and Wong, W.H. (1994). Sequential imputations and Bayesian missing data problems. *J. Am. Stat. Assoc.* 89: 278–288.
- Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* 36: 493–499.
- Krzanowski, W.J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: a hypothesis-testing approach. *Biometrics* 38: 991–1002.
- Kulldorff, G. (1961). *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*. Stockholm and New York: Almqvist and Wiksell and Wiley.
- Laird, N.M. and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics* 38: 963–974.
- Lange, K. (1995a). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. B* 57: 425–437.
- Lange, K. (1995b). A quasi-Newtonian acceleration of the EM algorithm. *Stat. Sin.* 5: 1–18.
- Lange, K., Little, R.J., and Taylor, J.M.G. (1989). Robust statistical inference using the t distribution. *J. Am. Stat. Assoc.* 84: 881–896.
- LaVange, L.M. (1983). The analysis of incomplete longitudinal data with modeled covariance matrices. In: *Mimeo 1449*. Institute of Statistics, University of North Carolina.
- Lazzeroni, L.C. and Little, R.J. (1998). Random-effects models for smoothing post-stratification weights. *J. Off. Stat.* 14 (1): 61–78.
- Ledolter, J. (1979). A recursive approach to parameter estimation in regression and time series problems. *Commun. Stat. – Theor. Methods A8*: 1227–1245.

- Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *J. R. Stat. Soc. B* 58: 619–678 (with discussion).
- Lee, Y. and Nelder, J.A. (2001). Hierarchical generalized linear models: a synthesis of generalised linear models, random effects models and structured dispersions. *Biometrika* 88: 987–1006.
- Lee, Y. and Nelder, J.A. (2009). Likelihood inference for models with unobservables: another view. *Statist. Sci.* 24 (3): 255–302 (with discussion).
- Lee, H., Rancourt, E., and Särndal, C.E. (2002). Variance estimation from survey data under single imputation. In: *Survey Nonresponse*, Chapter 21 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little). New York: Wiley.
- Lee, Y., Nelder, J.A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. London: Chapman and Hall.
- Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Am. Stat. Assoc.* 112: 1759–1769.
- Li, K.H., Meng, X.-L., Raghunathan, T.E., and Rubin, D.B. (1991a). Significance levels from repeated p -values with multiply-imputed data. *Stat. Sin.* 1: 65–92.
- Li, K.H., Raghunathan, T.E., and Rubin, D.B. (1991b). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *J. Am. Stat. Assoc.* 86: 1065–1073.
- Li, F., Baccini, M., Mealli, F., Zell, E.R., Frangakis, C., and Rubin, D.B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *J. Comput. Graph. Stat.* 23 (3): 877–892.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.
- Lillard, L., Smith, J.P., and Welch, F. (1982). *What Do We Really Know About Wages: The Importance of Nonreporting and Census Imputation*. Santa Monica, CA: The Rand Corporation.
- Lillard, L., Smith, J.P., and Welch, F. (1986). What do we really know about wages? The importance of nonreporting and census imputation. *J. Pol. Econ.* 94: 489–506.
- Lindley, D.V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2, Inference*. Cambridge: Cambridge University Press.
- Lipsitz, S.R., Ibrahim, J.G., and Zhao, L.P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Am. Stat. Assoc.* 94: 1147–1160.
- Little, R.J. (1976). Inference about means from incomplete multivariate data. *Biometrika* 63: 593–604.
- Little, R.J. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *J. R. Stat. Soc. B* 41: 76–87.
- Little, R.J. (1982). Models for nonresponse in sample surveys. *J. Am. Stat. Assoc.* 77: 237–250.

- Little, R.J. (1985a). A note about models for selectivity bias. *Econometrica* 53: 1469–1474.
- Little, R.J. (1985b). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bull. Int. Stat. Inst.* 15, 1: 1–15.
- Little, R.J. (1986). Survey nonresponse adjustments. *Int. Stat. Rev.* 54: 139–157.
- Little, R.J. (1988a). Small sample inference about means from bivariate normal data with missing values. *Comput. Stat. Data Anal.* 7: 161–178.
- Little, R.J. (1988b). Robust estimation of the mean and covariance matrix from data with missing values. *Appl. Stat.* 37: 23–38.
- Little, R.J.A. (1988c). Missing data in large surveys. *J. Bus. Econ. Stat.* 6: 287–301 (with discussion).
- Little, R.J. (1992). Regression with missing X's: a review. *J. Am. Stat. Assoc.* 87: 1227–1237.
- Little, R.J. (1993a). Statistical analysis of masked data. *J. Off. Stat.* 9: 407–426.
- Little, R.J. (1993b). Post-stratification: a modeler's perspective. *J. Am. Stat. Assoc.* 88: 1001–1012.
- Little, R.J. (1993c). Pattern-mixture models for multivariate incomplete data. *J. Am. Stat. Assoc.* 88: 125–134.
- Little, R.J. (1994). A class of pattern-mixture models for normal missing data. *Biometrika* 81 (3): 471–483.
- Little, R.J. (1995). Modeling the drop-out mechanism in longitudinal studies. *J. Am. Stat. Assoc.* 90: 1112–1121.
- Little, R.J. (1997). Biostatistical analysis with missing data. In: *Encyclopedia of Biostatistics* (ed. P. Armitage and T. Colton). London: Wiley.
- Little, R.J.A. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *Am. Stat.* 60 (3): 213–223.
- Little, R.J. (2008). Selection and pattern-mixture models. In: *Advances in Longitudinal Data Analysis*, Chapter 18 (ed. G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), 409–431. London: CRC Press.
- Little, R.J.A. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Stat. Sin.* 14: 949–968.
- Little, R.J. and Rubin, D.B. (1983a). Incomplete data. In: *Encyclopedia of Statistical Sciences*, vol. 4 (ed. S. Kotz), 46–53. Wiley.
- Little, R.J. and Rubin, D.B. (1983b). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *Am. Stat.* 37: 218–220.
- Little, R.J. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, 1e. New York: Wiley.
- Little, R.J. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2e. New York: Wiley.
- Little, R.J. and Schenker, N. (1994). Missing data. In: *Handbook for Statistical Modeling in the Social and Behavioral Sciences*, Chapter 2 (ed. G. Arminger, C.C. Clogg and M.E. Sobel), 39–75. New York: Plenum.

- Little, R.J. and Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 72: 497–512.
- Little, R.J.A. and Su, H.L. (1987). Missing-data adjustments for partially-scaled variables. In: *Proceedings of the Survey Research Methods Section, 1987*, 644–649. American Statistical Association.
- Little, R.J. and Su, H.L. (1989). Item nonresponse in panel surveys. In: *Panel Surveys* (ed. D. Kasprzyk, G. Duncan and M.P. Singh), 400–425. New York: Wiley.
- Little, R.J. and Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Stat. Med.* 22: 1589–1599.
- Little, R.J. and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Surv. Methods* 31: 161–168.
- Little, R.J. and Wang, Y.-X. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 52: 98–111.
- Little, R.J. and Yau, L. (1996). Intent-to-treat analysis in longitudinal studies with drop-outs. *Biometrics* 52: 1324–1333.
- Little, R.J. and Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Appl. Stat.* 60 (4): 591–605.
- Little, R.J., Liu, F., and Raghunathan, T. (2004). Statistical disclosure techniques based on multiple imputation. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (ed. A. Gelman and X.-L. Meng), 141–152. New York: Wiley.
- Little, R.J., Long, Q., and Lin, X. (2009). A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance. *Biometrics* 65 (2): 640–649.
- Little, R.J., D'Agostino, R., Cohen, M.L. et al. (2012). Special report: The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* 367 (14): 1355–1360.
- Little, R.J., Rubin, D.B., and Zanganeh, S.Z. (2016a). Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *J. Am. Stat. Assoc.* 112: 314–320.
- Little, R.J., Wang, J., Sun, X. et al. (2016b). The treatment of missing data in a large cardiovascular clinical outcomes study. *Clin. Trials* 13 (3): 344–351.
- Liu, C.H. (1995). Missing-data imputation using the multivariate t distribution. *J. Multivariate Anal.* 53: 139–158.
- Liu, C.H. (1996). Bayesian robust multivariate linear regression with incomplete data. *J. Am. Stat. Assoc.* 91: 1219–1227.
- Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
- Liu, C.H. (2005). Robit regression: a simple robust alternative to logistic and probit regression. In: *Applied Bayesian Modeling and Causal Inference from*

- Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (ed. A. Gelman and X.-L. Meng), 227–238. New York: Wiley.
- Liu, J.S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Am. Stat. Assoc.* 93: 1032–1044.
- Liu, C.H. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence. *Biometrika* 81: 633–648.
- Liu, C.H. and Rubin, D.B. (1996). Markov-normal analysis of iterative simulations before their convergence. *J. Econometrics* 75: 69–78.
- Liu, C.H. and Rubin, D.B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* 85: 673–688.
- Liu, C.H. and Rubin, D.B. (2002). Model-based analysis to improve the performance of iterative simulations. *Stat. Sin.* 12: 751–767.
- Liu, C.H., Rubin, D.B., and Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* 85: 755–770.
- Liublinska, V. and Rubin, D.B. (2012). Enhanced tipping-point displays. In: *Proceedings of the Section on Survey Research Methods, 2012*, 3861–3686. American Statistical Association.
- Lohr, S. (2010). *Sampling: Design and Analysis*, 2e. Boston, MA: Cengage Learning.
- Lord, F.M. (1955). Estimation of parameters from incomplete data. *J. Am. Stat. Assoc.* 50: 870–876.
- Louis, T.A. (1982). Finding the observed information when using the EM algorithm. *J. R. Stat. Soc. B* 44: 226–233.
- Madow, W.G. and Olkin, I. (1983). *Incomplete Data in Sample Surveys: Proceedings of the Symposium*, vol. 3. New York: Academic Press.
- Madow, W.G., Nisselson, H., and Olkin, I. (eds.) (1983a). *Incomplete Data in Sample Surveys: Report and Case Studies*, vol. 1. New York: Academic Press.
- Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983b). *Incomplete Data in Sample Surveys: Theory and Bibliographies*, vol. 2. New York: Academic Press.
- Manski, C.F. and Lerman, S.R. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* 45: 1977–1988.
- Marini, M.M., Olsen, A.R., and Rubin, D.B. (1980). Maximum-likelihood estimation in panel studies with missing data. *Sociol. Methodol.* 11: 314–357.
- Marker, D.A., Judkins, D.R., and Winglee, M. (2002). Large-scale imputation for complex surveys. In: *Survey Nonresponse*, Chapter 22 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little). New York: Wiley.
- Matthai, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhya* 2: 145–152.
- McCullagh, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. B* 42: 109–142.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2e. New York: CRC Press.

- McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.* 92: 162–170.
- McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.* 44: 98–130.
- McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.
- Mealli, F. and Rubin, D.B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* 102 (4): 995–1000. Correction in *Biometrika* 103 (2): 491.
- Mega, J.L., Braunwald, E., Wiviott, S.D. et al. (2012). Rivaroxaban in patients with a recent acute coronary syndrome. *N. Engl. J. Med.* 366: 9–19.
- Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Stat. Soc. B* 51: 127–138.
- Meinert, C.L. (1980). Toward more definitive clinical trials, controlled. *Clin. Trials* 1: 249–261.
- Meltzer, A., Goodman, C., Langwell, K. et al. (1980). Develop physician and physician extender data bases. *Final Report, G-155*. Silver Springs, MD: Applied Management Sciences, Inc.
- Meng, X.-L. (1995). Multiple imputation with uncongenial sources of input. *Stat. Sci.* 10: 538–73 (with discussion).
- Meng, X.L. (2002). A congenial overview and investigation of multiple imputation inferences under uncongeniality. In: *Survey Nonresponse*, Chapter 23 (ed. R. Groves, D. Dillman, J. Eltinge and R. Little). New York: Wiley.
- Meng, X.-L. (2009). Decoding the H-likelihood. Discussion of “likelihood inference for models with unobservables: another view” by Lee, Y. and Nelder, J.A. *Stat. Sci.* 24 (3): 280–293.
- Meng, X.-L. and Pedlow, S. (1992). EM: a bibliographic review with missing articles. In: *Proceedings of the Statistical Computing Section*, 24–27. American Statistical Association.
- Meng, X.-L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Stat. Assoc.* 86: 899–909.
- Meng, X.-L. and Rubin, D.B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 79: 103–111.
- Meng, X.-L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80: 267–278.
- Meng, X.-L. and Rubin, D.B. (1994). On the global and component-wise rates of convergence of the EM algorithm. *Linear Algebra Appl.* 199: 413–425.
- Meng, X.L. and Van Dyk, D. (1997). The EM algorithm – an old folk song sung to a fast new tune. *J. R. Stat. Soc. B* 59: 511–567 (with discussion).
- Meng, X.-L. and Wong, W.H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat. Sin.* 6: 831–860.

- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N. et al. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21: 1087–1091.
- Miller, R.G. (1974). The Jackknife – a review. *Biometrika* 61: 1–15.
- Mislevy, R.J., Beaton, A.E., Kaplan, B., and Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *J. Educ. Meas.* 29 (2): 133–161.
- Mori, M., Woolson, R.F., and Woodsworth, G.G. (1994). Slope estimation in the presence of informative censoring: modeling the number of observations as a geometric random variable. *Biometrics* 50: 39–50.
- Morrison, D.F. (1971). Expectations and variances of maximum likelihood estimates of the multivariate normal distribution parameters with missing data. *J. Am. Stat. Assoc.* 66: 602–604.
- Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.
- Murray, G.D. and Findlay, J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Stat. Med.* 7: 941–946.
- National Assessment of Educational Progress (2016). Overview of the NAEP Assessment Design. <https://nces.ed.gov/nationsreportcard/tdw/overview>.
- National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. Washington, DC: National Academy Press.
- Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc. A* 97: 558–606.
- Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *J. Stat. Softw.* 9: 1–54.
- Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., and Bent, D.H. (1975). *SPSS Statistical Package for the Social Sciences*, 2e. New York: McGraw-Hill.
- Nordheim, E.V. (1984). Inference from nonrandomly missing data: an example from a genetic study on Turner's Syndrome. *J. Am. Stat. Assoc.* 79: 772–780.
- Oh, H.L. and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. In: *Incomplete Data in Sample Surveys: Theory and Annotated Bibliography*, vol. 2 (ed. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press.
- Olkin, I. and Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.* 32: 448–465.
- Orchard, T. and Woodbury, M.A. (1972). A missing information principle: theory and applications. In: *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, 697–715.
- Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Stat. Med.* 12: 1723–1732.
- Pauli, F., Racugno, W., and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Stat. Sin.* 21: 149–164.

- Pearce, S.C. (1965). *Biological Statistics: An Introduction*. New York: McGraw-Hill.
- Pettitt, A.N. (1985). Re-weighted least squares estimation with censored and grouped data: an application of the EM algorithm. *J. R. Stat. Soc. B* 47: 253–261.
- Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. New York: Wiley.
- Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 51: 313–326.
- Preece, D.A. (1971). Iterative procedures for missing values in experiments. *Technometrics* 13: 743–753.
- Pregibon, D. (1977). Typical survey data: estimation and imputation. *Surv. Methodol.* 2: 70–102.
- Press, S.J. and Scott, A.J. (1976). Missing variables in Bayesian regression, II. *J. Am. Stat. Assoc.* 71: 366–369.
- Press, S.J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *J. Am. Stat. Assoc.* 73: 699–705.
- Raghunathan, T.E. (2015). *Missing Data Analysis in Practice*. New York: Chapman and Hall / CRC.
- Raghunathan, T.E. and Grizzle, J.E. (1995). A split questionnaire design. *J. Am. Stat. Assoc.* 90: 55–63.
- Raghunathan, T.E. and Rubin, D.B. (1998). Roles for Bayesian techniques in survey sampling. In: *Proceedings of the Silver Jubilee Meeting of the Statistical Society of Canada*, 51–55.
- Raghunathan, T., Lepkowski, J., Van Hoewyk, M., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* 27 (1): 85–95. For associated IVEWARE software see <http://www.isr.umich.edu/src/smp/ive/>.
- Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* 19: 1–16.
- Rao, C.R. (1965). *Linear Statistical Inference*. New York: Wiley.
- Rao, C.R. (1972). *Linear Statistical Inference and Its Applications*. New York: Wiley.
- Rao, J.N.K. (1996). On variance estimation with imputed survey data. *J. Am. Stat. Assoc.* 91: 499–506.
- Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 79: 811–822.
- Rässler, S. (2002). *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York: Springer.
- Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika* 95 (4): 933–946.
- Robins, J.M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* 90: 122–129.

- Robins, J.M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* 87(1): 113–124.
- Robins, J.M., Rotnitsky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 90: 106–121.
- Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.
- Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score. *Am. Stat.* 39: 33–38.
- Rotnitzky, A., Robins, J.M., and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Stat. Assoc.* 93: 1321–1339.
- Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* 59 (4): 829–836.
- Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. New York: Chapman and Hall / CRC.
- Rubin, D.B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Appl. Stat.* 21: 136–141.
- Rubin, D.B. (1973a). Matching to remove bias in observational studies. *Biometrics* 29: 159–183.
- Rubin, D.B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29: 185–203.
- Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *J. Am. Stat. Assoc.* 69: 467–474.
- Rubin, D.B. (1976a). Inference and missing data. *Biometrika* 63: 581–592 (with discussion).
- Rubin, D.B. (1976b). Non-iterative least squares estimates, standard errors and F-tests for any analysis of variance with missing data. *J. R. Stat. Soc. B* 38: 270–274.
- Rubin, D.B. (1976c). Comparing regressions when some predictor variables are missing. *Technometrics* 18: 201–206.
- Rubin, D.B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.* 72: 538–543.
- Rubin, D.B. (1978a). Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 7: 34–58.
- Rubin, D.B. (1978b). Multiple imputations in sample surveys. In: *Proceedings of the Survey Research Methods Section, 1978*, 20–34. American Statistical Association.
- Rubin, D.B. (1979). Illustrating the use of multiple imputation to handle nonresponse in sample surveys. *Proceedings of the 1979 International Statistical Institute - IASS*, Manila.

- Rubin, D.B. (1983a). Iteratively reweighted least squares. In: *Encyclopedia of Statistical Sciences*, vol. 4 (ed. S. Kotz, N.L. Johnson and C.B. Read), 272–275. New York: Wiley.
- Rubin, D.B. (1983b). Imputing income in the CPS. In: *The Measurement of Labor Cost* (ed. J. Triplett). Chicago: University of Chicago Press.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12: 1151–1172.
- Rubin, D.B. (1985a). The use of propensity scores in applied Bayesian inference. In: *Bayesian Statistics*, vol. 2 (ed. J.M. Bernardo, M.H. De Groot, D.V. Lindley and A.F.M. Smith), 463–472. Amsterdam: North-Holland.
- Rubin, D.B. (1985b). Comment on “A statistical model for positron emission tomography”. *J. Am. Stat. Assoc.* 80: 31–32.
- Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.* 4: 87–94.
- Rubin, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, D.B. (1987b). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. Discussion of Tanner and Wong (1987). *J. Am. Stat. Assoc.* 82: 543–546.
- Rubin, D.B. (1993). Satisfying confidentiality constraints through use of synthetic multiple-imputed microdata. *J. Off. Stat.* 9: 461–468.
- Rubin, D.B. (1994). Comment on “Missing data, imputation, and the bootstrap” by Bradley Efron. *J. Am. Stat. Assoc.* 89: 475–478.
- Rubin, D.B. (1996). Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* 91: 473–489 (with discussion).
- Rubin, D.B. (2000). The utility of counterfactuals for causal inference. Comment on A.P. Dawid, “causal inference without counterfactuals”. *J. Am. Stat. Assoc.* 95: 435–438.
- Rubin, D.B. (2002). Multiple imputation of NMES. *Proceedings of the International Conference on Quality in Official Statistics*, Stockholm, Sweden (14–15 May 2001).
- Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*, Wiley Classics Library Edition. New York: Wiley.
- Rubin, D.B. (2017). Commentary. *Stat. J. Int. Assoc. Off. Stat.* 33 (1): 239–240.
- Rubin, D.B. (2019). *Conditional Calibration and the Sage Statistician Survey Methodology*, June 2019, in press.
- Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Stat. Assoc.* 81: 366–374.
- Rubin, D.B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: a case study using agriculture industry codes. *J. Off. Stat.* 3: 375–387.

- Rubin, D.B. and Szatrowski, T.H. (1982). Finding maximum likelihood estimates for patterned covariance matrices by the EM algorithm. *Biometrika* 69: 657–660.
- Rubin, D.B. and Thayer, D. (1978). Relating tests given to different samples. *Psychometrika* 43: 3–10.
- Rubin, D.B. and Thayer, D.T. (1982). EM algorithms for factor analysis. *Psychometrika* 47: 69–76.
- Rubin, D.B. and Thayer, D.T. (1983). More on EM for ML factor analysis. *Psychometrika* 48: 253–257.
- Rubin, D.B. and Thomas, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Stat.* 20: 1079–1093.
- Rubin, D.B. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Stat. Assoc.* 95: 573–585.
- Rubin, D.B., Stern, H., and Vehovar, V. (1996). Handling ‘don’t know’ survey responses: the case of the Slovenian plebiscite. *J. Am. Stat. Assoc.* 90: 822–828.
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press.
- Sande, I.G. (1983). Hot deck imputation procedures. In: *Incomplete Data in Sample Surveys: Symposium on Incomplete Data, Proceedings*, vol. 3 (ed. W.G. Madow and I. Olkin). New York: Academic Press.
- SAS (1992). The Mixed Procedure, Chapter 16 in SAS/STAT Software: Changes and Enhancements, Release 6.07. *Technical Report P-229*. Cary, NC: SAS Institute, Inc.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. New York: CRC Press.
- Schafer, J.L. (1998). Multiple imputation: a primer. *Stat. Methods Med. Res.* 8: 3–15.
- Scharfstein, D., Rotnitsky, A., and Robins, J. (1999). Adjusting for nonignorable dropout using semiparametric models. *J. Am. Stat. Assoc.* 94: 1096–1146 (with discussion).
- Schenker, N. and Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Comput. Stat. Data Anal.* 22: 425–446.
- Schieber, S.J. (1978). A comparison of three alternative techniques for allocating unreported social security income on the survey of the low-income aged and disabled. In: *Proceedings of the Survey Research Methods Section, 1978*, 212–218. American Statistical Association.
- Schluchter, M.D. (1992). Methods for the analysis of informatively censored longitudinal data. *Stat. Med.* 11 (14–15): 1861–1870.
- Schluchter, M.D. and Jackson, K.L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *J. Am. Stat. Assoc.* 84: 42–52.
- Seaman, S.R. and White, I.R. (2011). Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* 22: 278–295.

- Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by “missing at random?”. *Stat. Sci.* 28 (2): 257–268.
- Shao, J. (2002). Replication methods for variance estimation in complex surveys with imputed data. In: *Survey Nonresponse*, Chapter 20 (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J. Little), 303–314. New York: Wiley.
- Shao, J., Chen, Y., and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *J. Am. Stat. Assoc.* 93: 819–831.
- Shih, W.J., Quan, H., and Chang, M.N. (1994). Estimation of the mean when data contain non-ignorable missing values from a random effects model. *Stat. Probabil. Lett.* 19: 249–257.
- Shumway, R.H. and Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* 3: 253–264.
- Sinha, D. and Ibrahim, J.G. (2003). A Bayesian justification of Cox’s partial likelihood. *Biometrika* 90 (3): 629–641.
- Skinner, C.J., Smith, T.M.F., and Holt, D. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *Am. Stat.* 46: 84–88.
- Snedecor, G.W. and Cochran, W.G. (1967). *Statistical Methods*. Ames, IA: Iowa State University Press.
- Spiegelman, D., Carroll, R.J., and Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Stat. Med.* 20: 139–160.
- Stuart, A. and Ord, J.K. (1994). *Kendall’s Advanced Theory of Statistics: Distribution Theory*, 6e, vol. 1. New York: Arnold.
- Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Stat.* 1: 49–58.
- Szatrowski, T.H. (1978). Explicit solutions, one iteration convergence and averaging in the multivariate normal estimation problem for patterned means and covariances. *Ann. Inst. Stat. Math.* 30: 81–88.
- Szpiro, A., Rice, K.M., and Lumley, T. (2010). Model-robust regression and a Bayesian “sandwich” estimator. *Ann. Appl. Stat.* 4 (4): 2099–2113.
- Tang, G., Little, R.J., and Raghunathan, T. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* 90: 747–764.
- Tanner, M.A. (1996). *Tools for Statistical Inference*, 3e. New York: Springer.
- Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82: 528–550 (with discussion).
- Ten Have, T.R., Pulkstenis, E., Kunselman, A., and Landis, J.R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics* 54: 367–383.
- Ten Have, T.R., Reboussin, B.A., Miller, M.E., and Kunselman, A. (2002). Mixed effects logistic regression models for multiple longitudinal binary functional

- limitation responses with informative drop-out and confounding by baseline outcomes. *Biometrics* 58 (1): 137–144.
- Thisted, R.A. (1988). *Elements of Statistical Computing*. New York: CRC Press.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.
- Tocher, K.D. (1952). The design and analysis of block experiments. *J. R. Stat. Soc. B* 14: 45–100.
- Trawinski, I.M. and Bargmann, R.W. (1964). Maximum likelihood with incomplete multivariate data. *Ann. Math. Stat.* 35: 647–657.
- Tu, X.M., Meng, X.-L., and Pagano, M. (1993). The AIDS epidemic: estimating survival after AIDS diagnosis from surveillance data. *J. Am. Stat. Assoc.* 88: 26–36.
- Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. New York: Springer.
- Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *J. Off. Stat.* 20 (1): 1–18.
- Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. New York: Chapman Hall / CRC.
- Van Buuren, S. and Oudshoorn, C.G.M (1999). Flexible Multivariate Imputation by MICE. *Report TNO/VGZ/PG 99.054*. Leiden: TNO Preventie en Gezondheid. For associated software see <https://stefvanbuuren.name/mice>.
- Van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Stat. Med.* 18: 681–694.
- Van Dyk, D.A., Meng, X.L., and Rubin, D.B. (1995). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Stat. Sin.* 5: 55–75.
- Van Praag, B.M.S., Dijkstra, T.K., and Van Velzen, J. (1985). Least-squares theory based on general distributional assumptions with an application to the incomplete observations problem. *Psychometrika* 50: 25–36.
- Vardi, Y., Shepp, L.A., and Kaufman, L. (1985). A statistical model for positron emission tomography. *J. Am. Stat. Assoc.* 80: 8–37.
- Ventura, L., Cabras, S., and Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Am. Stat. Assoc.* 104: 768–774.
- Von Hippel, P.T. (2007). Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociol. Methodol.* 37: 83–117.
- Von Neumann, J. (1951). Various techniques used in connection with random digits. *Nat Bur. Stand. Appl. Math. Ser.* 12: 36–38.
- Wachter, K.W. and Trussell, J. (1982). Estimating historical heights. *J. Am. Stat. Assoc.* 77: 279–301.
- Ware, J.H. (1985). Linear models for the analysis of longitudinal studies. *Am. Stat.* 39: 95–101.

- Weisberg, S. (1980). *Applied Linear Regression*. New York: Wiley.
- West, B. and Little, R.J. (2013). Nonresponse adjustment of survey estimates based on auxiliary variables subject to error. *Appl. Stat.* 62 (2): 213–231.
- White, H. (1982). Maximum likelihood under misspecified models. *Econometrica* 50: 1–25.
- Wilkinson, G.N. (1958a). Estimation of missing values for the analysis of incomplete data. *Biometrics* 14: 257–286.
- Wilkinson, G.N. (1958b). The analysis of variance and derivation of standard errors for incomplete data. *Biometrics* 14: 360–384.
- Wilks, S.S. (1932). Moments and distribution of estimates of population parameters from fragmentary samples. *Ann. Math. Stat.* 3: 163–195.
- Wilks, S.S. (1963). *Mathematical Statistics*. New York: Wiley.
- Winer, B.J. (1962). *Statistical Principles in Experimental Design*. New York: McGraw-Hill.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- Woolson, R.F. and Clarke, W.R. (1984). Analysis of categorical incomplete longitudinal data. *J. R. Stat. Soc. A* 147: 87–99.
- Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* 11: 95–103.
- Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44: 175–188.
- Wu, C.F.J. and Hamada, M. (2009). *Experiments: Planning, Analysis, and Optimization*, 2e. New York: Wiley.
- Xie, X. and Meng, X.-L. (2017). Dissecting multiple imputation from a multiphase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Stat. Sin.* 27: 1485–1594.
- Yang, Y. and Little, R. (2015). A comparison of doubly robust estimators of the mean with missing data. *J. Stat. Comput. Simul.* 85 (16): 3383–3403.
- Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.* 1: 129–142.
- Yuan, Y. and Little, R.J. (2009). Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics* 65 (2): 478–486.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Am. Stat. Assoc.* 57: 348–368.
- Zhang, G. and Little, R.J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* 65: 911–918.
- Zhang, G. and Little, R.J. (2011). A comparative study of doubly robust estimators of the mean with missing data. *J. Stat. Comput. Simul.* 81 (12): 2039–2058.
- Zhao, L.P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Stat. Med.* 11: 769–782.

Author Index

a

- Aboyomi, K. 243
 Afifi, A.A. 3, 47
 Airoldi, E.M. 353
 Aitkin, M. 145, 345
 Albert, P.S. 384
 Allan, F.G. 34, 44
 Amemiya, T. 358
 An, H. 21, 298, 300
 Anderson, R.L. 30
 Anderson, T.W. 113, 153, 175, 176,
 341
 Andrews, D.F. 285
 Andridge, R.H. 68, 80, 378
 Angrist, I.D. 7
 Azen, S. 63, 66

b

- Baccini, M. 243
 Bailar, B.A. 78, 84
 Bailar, J.C. 78
 Bailey, L. 84
 Baker, S.G. 385, 388, 389
 Bang, H. 299
 Bard, Y. 142
 Bargmann, R.W. 252
 Barnard, J. 7, 98, 235
 Bartlett, M.S. 35
 Bates, D.M. 269, 299
 Baum, L.E. 188

- Beale, E.M.L. 188, 251, 252, 283
 Beaton, A.E. 19, 166
 Becker, M.P. 187
 Bentler, P.M. 283
 Besag, J. 201
 Bethlehem, J.G. 59
 Bickel, P.J. 285
 Bishop, Y.M.M. 60, 203, 302, 319–320
 Bondarenko, I. 243
 Boshuizen, H.C. 395
 Box, G.E.P. 29, 125, 159, 182, 273, 369
 Box, M.J. 142
 Braunwald, E. 398
 Breslow, N.E. 145
 Brown, C.H. 374
 Brownlee, K.A. 264
 Buck, S.F. 72, 82
 Burge, H. 19

c

- Cabras, S. 140
 Caines, P.E. 276
 Carlin, J.B. 59, 109, 113, 119, 120, 123,
 127, 129–131, 150, 159, 182,
 231–232, 322, 323
 Carpenter, J.R. 3, 96, 364
 Carroll, R.J. 19, 75, 280, 384
 Cassel, C.M. 57, 65
 Chang, M.N. 384
 Chaurasia, A. 236

Chen, M.-H. 344

Chen, R. 231

Chen, T. 315

Chen, Y. 86

Chew, G. 19

Chu, H. 19

Church, T.R. 19

Clarke, W.R. 4, 6

Clayton, D.G. 145

Cochran, W.G. 29, 39, 51, 77, 79, 155,
262

Cohen, M.L. 398

Cole, S.R. 19

Corby, C. 84

Cox, D.R. 109, 119, 369, 380

Cox, G. 29, 39

Crainiceanu, C.M. 19, 280, 299

Curry, J. 63, 66

Czajka, J.L. 56

d

D'Agostino, R. 398

David, M.H. 68, 69, 70, 78, 371

Davidian, M. 19

Davies, O.L. 29

Day, N.E. 345

DeGroot, M.H. 119, 142

DeGruttola, V. 384

Dempster, A.P. 67, 166, 186, 188, 190,
194, 196–198, 252, 262, 264, 286

Diggle, P. 383

Dijkstra, T.K. 63

Dillman, D. 10

Ding, P. 52

Dodge, Y. 32, 44

Draper, N.R. 32, 142, 172

Du, I. 7

Dunson, D.B. 318

e

Edwards, A.W.F. 118

Efron, B. 90–92, 122–123, 219, 220

Eilers, P.H.C. 298

Ekholm, A. 5

Elashoff, R.M. 3, 47

Eltinge, J. 10

Ernst, L.R. 68

Ezzati-Rice, T. 101, 245

f

Fay, R.E. 86, 94, 95, 100, 101, 244, 385

Fienberg, S.E. 60, 203, 302, 315, 319,
320

Findlay, J.G. 22

Firth, D. 116

Follman, D.A. 384

Ford, B.N. 68

Frangakis, C.E. 7, 384

Franks, A.M. 353

Freedman, L.S. 19

Frumento, P. 118, 126, 342

Fuchs, C. 308, 310, 324, 326

Fuller, W.A. 19

g

Galati, J. 150

Galke, W. 358

Gelman, A.E. 59, 109, 113, 119, 120,
123, 127, 129–131, 159, 182, 230,
232, 243, 264, 322–323

Geyer, C.J. 230

Giusti, C. 391–394, 396, 397

Glynn, R.J. 49, 352, 353, 365

Goel, K. 142

Goodman, C. 277, 278

Goodman, L.A. 301, 319

Goodnight, J.H. 301

Goodrich, R.L. 276

Greenland, S. 19

Greenlees, W.S. 362, 364

Grizzle, J.E. 19

Groves, R. 10

Guo, Y. 19, 180, 182, 277, 279

Gupta, N.K. 276

h

- Haberman, S.J. 319
 Haitovsky, Y. 63
 Hajek, J. 24, 52
 Hamada, M. 188
 Hampel, F.R. 285
 Hansen, M.H. 51
 Hanson, R.H. 78
 Harel, O. 236
 Hartley, H.O. 3, 34, 117, 188, 252, 314
 Harvey, A.C. 273
 Harville, D.A. 117, 118
 Hasselblad, V. 358
 Hastings, W.K. 232
 Healy, M.J.R. 34
 Heckman, J. 362, 369, 403
 Heeringa, S.G. 359, 360
 Heitjan, D.F. 27, 145, 147, 299,
 Henderson, C.R. 145
 Herzog, T. 68, 96, 97, 99
 Higgins, K.M. 19
 Hill, I. 7
 Hills, S.E. 264
 Hinkley, D.V. 109, 119, 122
 Hirabayashi, S.M. 56
 Hirano, K. 7
 Hocking, R.R. 3, 252, 309
 Holland, P.W. 60, 203, 260, 302, 319, 320, 353
 Holt, D. 59, 101
 Horvitz, D.G. 24, 53
 Huber, P.J. 119
 Hull, C.H. 24
 Hunter, J.S. 29
 Hunter, W.G. 29, 142
 Hurwitz, W.N. 51

i

- Ibrahim, J.G. 58, 140, 344
 Imbens, G.W. 7
 Ireland, C.T. 60

j

- Jackson, D. 150
 Jacobsen, M. 147
 Jamshidian, M. 187, 205, 209, 210
 Janssens, W. 371
 Jarrett, R.G. 30, 35, 42
 Jenkins, G.M. 273
 Jenkins, J.G. 24
 Jennrich, R.I. 187, 201, 205, 209, 210, 269–272
 Johnson, W. 101, 245
 Jones, R.H. 273
 Judkins, D.R. 68, 79, 80
 Jurek, A.M. 19
- k**
- Kalman, R.E. 273
 Kalton, G. 68, 78
 Kang, J.D.Y. 300
 Kaplan, B. 19
 Kaufman, L. 316
 Keiding, N. 147
 Kempthorne, O. 29
 Kennickell, A.B. 243
 Kent, J.T. 207
 Kenward, M.G. 3, 96, 364, 383
 Khare, M. 101, 245
 Kim, J.O. 63, 66
 Kipnis, V. 19, 75
 Kish, L. 54, 68, 78
 Kleinbaum, D.G. 49
 Knook, D.L. 395
 Kong, A. 231
 Krishnan, T. 187, 208
 Krzanowski, W.J. 340, 348
 Kullback, S. 60
 Kulldorff, G. 355
 Kunselman, A. 384
 Kupper, L.L. 49

I

- Laird, N.M. 49, 186, 188, 190, 194, 196–198, 252, 262, 264, 269, 286, 352, 365, 385, 388, 389
 Landis, J.R. 384
 Lange, K. 187, 209, 294
 Langwell, K. 277, 278
 LaVange, L.M. 294–295
 Lazzeroni, L.C. 59
 Ledolter, J. 276
 Lee, H. 86
 Lee, Y. 142, 144–145
 Lepkowski, J. 243, 392
 Lerman, S.R. 58
 Levy, M. 243
 Li, F. 243
 Li, K.H. 236–237
 Li, X. 52
 Liang, K.-Y. 57
 Lillard, L. 369–371
 Lin, X. 145
 Lindley, D.V. 159
 Lipsitz, S. 58
 Lipsitz, S.R. 58, 344
 Little, R.J. 3–4, 19, 21, 49, 53, 55–56, 59–60, 65, 68–70, 75, 78, 82, 84, 132, 139, 150, 159, 161, 180, 182, 183, 188, 219, 239, 241, 251, 261, 268, 277, 279, 290–291, 294, 298, 300, 315, 329, 332–333, 342–343, 348, 352–353, 359–360, 362, 371, 374, 375, 378–383, 385, 389, 391–394, 396, 398–401, 403–404
 Liu, C.H. 205–206, 231–232, 261, 293, 329, 347
 Liu, J.S. 231
 Liublinska, V. 400, 402–403
 Lohr, S. 51
 Lord, F.M. 176
 Louis, T.A. 199, 208, 220
 Lumley, T. 123

m

- Madow, W.G. 10, 51
 Maldonado, G. 19
 Manski, C.F. 58
 Marini, M.M. 11, 163
 Marker, D.A. 68, 79, 80
 Marx, B.D. 298
 Matthai, A. 62
 McConnell, D.S. 19, 180, 182, 279
 McCullagh, P. 116, 301
 McCulloch, C.E. 145
 McKendrick, A.G. 188
 McLachlan, G.J. 187, 208
 Mealli, F. 23, 118, 126, 132, 243, 342
 Mega, J.L. 398
 Mehra, R.K. 276
 Meilijson, I. 208, 221
 Meinert, C.L. 13
 Meltzer, A. 277–278
 Meng, X.L. 145, 187, 199–201, 206, 214, 216, 222, 232, 235–237, 244, 321
 Metropolis, N. 232
 Midthune, D. 19
 Miller, M.E. 384
 Miller, R.G. 92, 123
 Mislevy, R.J. 19
 Morgenstern, H. 49
 Mori, M. 384
 Morrison, D.F. 182
 Muirhead, R.J. 132
 Murray, G.D. 22

n

- Nelder, J.A. 116, 142, 144–145
 Neyman, J. 89
 Ngo, L. 299
 Nie, N.H. 24
 Nisselson, H. 10
 Nordheim, E.V. 385

o

- Oh, H.L. 54, 55, 65, 70
 Olkin, I. 10, 330

- Olsen, A.R. 11, 163
 Orchard, T. 3, 188, 199, 252
 Ord, J.K. 114
 Oudshoorn, C.G.M. 395
 Oxspring, H.H. 309, 311
- p**
 Pacini, B. 118, 126, 342
 Park, T. 58
 Pauli, F. 140
 Pawitan, Y. 145
 Pearce, S.C. 35
 Pedlow, S. 187
 Petrie, T. 188
 Pettitt, A.N. 288
 Phillips, G.D.A. 273
 Pinheiro, J.C. 269, 299
 Pocock, S.J. 80
 Potthoff, R.F. 270
 Preece, D.A. 35
 Pregibon, D. 385
 Press, S.J. 142, 330, 348
 Pulkstenis, E. 384
- q**
 Quan, H. 384
- r**
 Rässler, S. 177
 Racine-Poon, A. 264
 Racugno, W. 140
 Raghunathan, T.E. 3, 19, 21, 236–237, 243, 245, 359–360, 378, 392
 Rancourt, E. 86
 Rao, C.R. 43, 113, 119, 194
 Rao, J.N.K. 86, 94, 95, 100, 117
 Reboussin, B.A. 384
 Reece, J.S. 362, 364
 Reiter, J.P. 21, 281
 Rice, K.M. 123
 Rinke de Wit, T.F. 371
 Robins, J.M. 27, 58, 65, 244, 299
- Ronchetti, E.M. 285
 Rosenbaum, P.R. 56, 299
 Rosenbluth, A.W. 232
 Rosenbluth, M.N. 232
 Rotnitzky, A. 27, 58, 65, 299
 Rousseeuw, P.J. 285
 Roy, J. 384
 Roy, S.N. 270
 Royall, R. 118
 Rubin, D.B. 3, 4, 7, 8, 10–11, 13, 21, 23, 27, 38–39, 42, 51, 56, 67–68, 79, 82, 92, 96, 98–100, 118, 125–126, 132, 139, 145, 150, 163, 172, 175, 177–178, 180, 183, 186, 188, 199–202, 205–206, 214, 216, 222, 230–232, 236–237, 239, 240, 243–244, 245, 252, 261–262, 264, 286, 288–289, 316, 323, 329, 342, 345, 352, 366–367, 371, 389
 Ruppert, D. 19, 280, 299
- s**
 Samuhel, M.E. 68, 69, 70, 78, 371
 Sande, I.G. 68
 Särndal, C.E. 57, 65, 86
 Schafer, J.L. 3, 269, 300, 323
 Scharfstein, D.O. 27, 58
 Schenker, N. 3, 96, 98, 100, 235, 241, 244
 Scheuren, F.S. 54–55, 65, 70
 Schieber, S.J. 69
 Schluchter, M.D. 117, 201, 269–270, 329, 332–333, 342–343, 348
 Scott, A.J. 142
 Seaman, S.R. 27, 150
 Shao, J. 86, 94, 95
 Sheehan, K.M. 19
 Shepp, L.A. 316
 Shih, W.J. 384
 Shumway, R.H. 273, 276
 Sinha, D. 140
 Skinner, C.J. 5, 101
 Smith, A.F.M. 232, 264

- Smith, H. 32, 173
Smith, J.P. 369–371
Smith, T.M.F. 59, 101
Snedecor, G.W. 262
Solenberger, P. 243, 392
Soules, G. 188
Spiegelman, D. 75
Stahel, W.A. 285
Stead, A.G. 358
Stefanski, L.A. 19, 75, 280
Stern, H.S. 8, 23, 109, 113, 119, 120,
123, 127, 129–131, 159, 182, 231,
232, 322, 389
Stoffer, D.S. 273, 276
Stuart, A. 114
Su, H.L. 80, 359
Suh, E.B. 384
Sun, X. 398–401
Sundberg, R. 188
Szatrowski, T.H. 202, 259
Szapiro, A. 123
- t**
- Tanaka, J.S. 283
Tang, G. 378
Tanner, M.A. 224, 293
Tanović, Z. 371
Tate, R.F. 330
Taylor, J.M.G. 241, 294
Ten Have, T.R. 384
Thayer, D. 126, 177–178, 180,
261
Thisted, R.A. 201
Thomas, N. 79
Thompson, D.J. 24, 53
Tiao, G.C. 125, 159, 182
Tibshirani, R. 123
Tobin, J. 358
Tocher, K.D. 42
Trawinski, I.M. 252
Triest, R.K. 68, 69, 70, 78,
371
Trussell, J. 17
- Tsutakawa, R.K. 262, 264
Tu, X.M. 384
Tyler, D.E. 207
- v**
- Vach, W. 346
Valliant, R. 61
van Buuren, S. 3, 96, 395
van der Gaag, J. 371
van Dyk, D. 206, 22
van Guilder, M. 63, 66
van Hoewyk, M. 243, 392
van Praag, B.M.S. 63
van Velzen, J. 63
Vardi, Y. 207, 316
Vartivarian, S. 53, 55
Vehovar, V. 8, 23, 389
Ventura, L. 140
von Hippel, P.T. 382
von Neumann, J. 232
- w**
- Wachter, K.W. 17
Wand, M.P. 299
Wang, J. 398–401
Wang, N. 244
Wang, S.A. 384
Wang, Y.-X. 378
Ware, J.H. 117, 264, 269
Weisberg, S. 32
Weiss, N. 188
Welch, F. 369–371
West, B. 375
Westmacott, M. 34
White, H. 119
White, I.R. 27
Wightman, L.E. 260
Wilkinson, G.N. 34, 42
Wilks, S.S. 62
Wilson, G.T. 345
Wilson, S. 330, 348
Winer, B.J. 29
Winglee, M. 68, 79, 80

- Wishart, J. 34, 44
Wiviot, S.D. 398
Wolter, K.M. 361
Wong, W.H. 224, 231, 232
Woodbury, M.A. 188, 199,
 252
Woodsworth, G.G. 384
Woolson, R.F. 4, 6, 384
Wretman, J.H. 57, 65
Wu, C.F.J. 29, 188
Wu, M.C. 384
Wu, Y. 206
- X**
- Xie, X. 244
Xing, C. 318

- y**
- Yang, I. 187
Yang, Y. 300
Yates, F. 33
Yau, L. 84
Yuan, Y. 385
- z**
- Zanganeh, S.Z. 139
Zeger, S.L. 57
Zellner, A. 201
Zhang, G. 298–300, 379–380
Zhang, N. 379–381
Zhao, L.P. 27, 58, 65, 299
Zhou, X.H. 7
Zieschang, K.D. 362, 364

Subject Index

a

- Acceleration techniques 35, 208–210
- Accelerated EM algorithm (AEM) 209. *See also* PXEM algorithm
- Acquired immune deficiency syndrome (AIDS), demographic survey 371–372
- Adaptive robust estimation 204–205, 291–292
- Adjusting ANOVA sums of squares and standard errors for filled-in missing values 42
- Adjustment cells 78, 85, 244
- AECM algorithm 205–206
- Algorithms, iterative 209. *See also* EM algorithm; extensions of EM algorithm; Newton–Raphson algorithm; Scoring algorithm
- Allocation, *see* Imputation
- Analysis of covariance (ANCOVA) method for missing data in experiments 35–45
- Analysis of variance (ANOVA) with missingness outcomes 29–45, 264–266
 - mixed-effects models 31, 262
 - one-way ANOVA with missing values 138–140
 - random-effects models 31, 138–140, 262–264

- Approximate covariance matrix 267
 - See also* Asymptotic covariance matrix of parameters or estimates
- Approximations using test statistics 235–238
- AR1 Model 275
- Asymptotic covariance matrix of parameters or estimates 119–120, 213
 - for categorical data 312–313
 - for general missing data pattern using Louis's method 220–221
 - for multiple regression 266–269
 - for multivariate normal data 252–253
 - sandwich estimator 123
- Asymptotic inferences 61
- Asymptotic standard errors, *see* Asymptotic covariance matrix of parameters or estimates
- Augmented covariance matrix, definition 168–169
- Autoregressive model 273–276
 - moving-average (ARMA) models 273
- Available-case analysis, weighted 50–54
 - comparisons with complete-case analysis 62–63

- Available-case analysis (*Continued*)
 comparisons with ML and data augmentation analysis 255
- b**
- Balanced incomplete block design 30
 Balanced repeated replication 61
 Banded covariance structure 273
 Bartlett's method 35–37
 Bayesian bootstrap, *see* Bootstrap, approximate Bayesian
 Bayesian approach 366–378
 Bayesian analysis, draws 373
 Bayesian inference 118–119, 126–132, 159–160, 174–175, 180, 223–246. *See also* Multiple imputation
 by direct simulation, for complete data 130–132
 for estimation with complete data 118–119
 for multinomial sample 129, 131
 for special patterns of missing data 159–160, 174–175, 180
 for specific missing-data problems contaminated normal model 288
 contingency tables 305, 315–316
 logistic regression 346–347
 log-linear models for contingency tables 321–322
 mixed continuous and categorical data 335–337, 341
 multinomial data 305, 315–316
 multiple linear regression 128–129, 130–131, 268
 multivariate interval-censored (coarsened) data 358–361
 multivariate linear regression 267–268
 multivariate normal sample 129, 131–132, 254
 multivariate normal regression 228–229
 multivariate t model 292–293
 random-effects models 264
 robust MANOVA 293–294
 univariate t sample 229
 for univariate normal sample 127–128
 iterative simulation methods 223–232
 assessing convergence 230–231
 bridge and path sampling 232
 data augmentation 223–226
 parameter-expanded 229
 Gibbs' sampler 226–229
 Metropolis–Hastings algorithm 232
 Sampling importance resampling (SIR) algorithm 232
 large-sample theory with complete data 119–121
 posterior standard errors 221
 simulating draws from posterior distribution 130–132
 for monotone bivariate normal sample 159–160, 182
 for one-to-one functions of parameters 130
 with complete data 126–132
 with incomplete data 132–141
 Bayesian iterative proportional fitting 321–322, 341
 Beta distribution as prior distribution for binomial 131, 225
 Between-imputation variance 97, 234
 Bias due to nonresponse. *See also* Consistent estimates from incomplete data; Nonignorable missing-data mechanism
 Binomial distribution 111, 211
 Bioassay data 19, 280
 Bivariate normal data, general pattern Bayes' inference by DA 224–225
 EM algorithm for ML estimates 191–193

- inference by multiple imputation 235
- Bivariate normal monotone pattern 153–161, 169–170
- Bayes inference 160–161
- ML estimation 154
- via SWEEP operator 169–170
- large-sample covariance matrix 157–158
- using SEM algorithm 216–217
- precision of estimation 157–161
- SAS statistical software 269
- Bootstrap 219–220
- approximate Bayesian 244
 - standard errors 90–92
 - for complete data 90–91
 - for imputed data 91
 - for ML estimates 240
- Box–Jenkins time series models 273
- Bracketed response formats 359
- Buck's method 72
- c**
- Calibrated Bayes inference 4
- Calibration 19–20, 374. *See also* response error models; Internal and external
- Canonical correlation 249
- Cauchy distribution 147, 149
- Censored survival data 16
- exponential sample 110, 355–356
 - with known censoring points 16
 - with stochastic censoring points 16
- Censoring mechanisms 147–148
- Census Bureau hot deck 97
- Census Bureau imputations 371
- Central limit theorem 52, 122
- Chained-equation multiple imputation, sensitivity analysis for 241–243
- Chi-squared distribution 123–124, 127, 131, 159–160
- Chi-squared statistics, *see* Goodness-of-fit statistics
- Cholesky factor 132
- Classified data, column distributions of the fully 388
- Coarsened at random (CAR) 355
- Coarsened data 17, 145–150
- likelihood theory for 145
- Cold-deck imputation 69
- Complete-data likelihood 225, 302
- Compound symmetry 270
- Computational strategies, alternative 185–187
- Conditional mean imputation 62–64
- Conjugate prior distribution 127, 129
- Contrasts in analysis of variance 44
- Convergence 60
- Correlations, estimates from incomplete data, inestimable 177
- Counted data, *see* Categorical data
- Covariance components models 262
- Current Population Survey (CPS) 78
- d**
- Darwin's data 345
- Data
- grouped 355
 - rounded 355
- Data augmentation (DA) algorithm 223–226
- Data coarsening 145
- Data matrix 3
- Degrees of freedom, correction for
- ANOVA 41
 - correction for covariance matrix 251
- see also* Likelihood ratio statistic
- Density function 109
- Dependent variables, missingness values in 264
- Design weights 24
- Dirichlet distribution 129, 306

- Discarding data 355
 Disclosure limitation 20–21
 Discriminant analysis 25, 184
 Distinct parameters 162
 Donor for imputation 76
 “Don’t know” stratum 390
 Double sampling 18
 Drop-outs
 outcome-dependent 383
 random-coefficient dependent 383–384
 treatment drop-outs versus analysis drop-outs 13–14
 Dummy variable regression 162
- e**
 EA’s, *see* Enumeration areas (EA’s)
 ECM algorithm 200–205
 for specific missing-data problems:
 loglinear models for contingency tables 203–204
 multivariate normal regression model 201–203
 univariate t sample with unknown degree of freedom 204–205
 ECME algorithm 205–206
 for univariate t sample with unknown degree of freedom 205–206
 Educational testing, examples 177–180
 EM algorithm 187–188. *See also* Maximum likelihood
 estimation convergence 193–194
 expectation step (E step) 188–193
 extensions of 200–205
 for exponential families 196–198
 maximization step (M step) 188–193
 rate of convergence 198–200
 standard errors from EM computations, *see* SEM algorithm theory 193–196
- Enumeration areas (EA’s) 87
 E step (expectation step) 188–193.
 See also EM algorithm
 Estimating equations 57
 Expectation conditional-maximization algorithm, *see* ECM algorithm
 Expectation-maximization algorithm, *see* EM algorithm
 Exploratory factor analysis 261
 Exponential data 110
 with censored values 144
 with grouped values 355–356
 Exponential sample 110
 Exponential family, EM for 196–198
- f**
 F distribution 31
 Factor analysis 261–262, 283
 with missing data 262, 283
 Factored likelihood 302–313
 for bivariate normal data 153–156
 for mixed continuous and categorical data 343–344
 for monotone pattern 162–166
 for multinomial data with monotone pattern 302–312
 precision of estimation 312–313
 for multivariate normal monotone data 302–308
 Bayes computations 305–308
 ML computations 305–308
 for nonmonotone patterns 175–184
 for partially-classified contingency tables 302–312
 for special nonmonotone patterns 175–179
 F distribution, Snedecor’s 31–32
 File matching, with two sets of variables 12
 Filling in for missing values, *see* Imputation
 Finite population 50
 correction 54

- randomization-based
inference 50–53
with nonresponse 54
- Follow-ups, decreased sensitivity of
inference 365
- Forecasting procedures 276
- Fraction of missing information 102,
198
- Fully missing variables 258–259
- g**
- Gamma distribution 131
- Gauss–Seidel algorithm 201
- Generalized estimating equations
(GEE) 27, 57, 287
- Generalized linear mixed models
142
- Generalized linear models 115–116
with missing covariates 344
- General location model 329–331
extensions with t distributions for
continuous variables 346–347
- ML estimation with missing values
331–337
- with parameter constraints
337–344
- Gibbs' sampler 226–229
monitoring convergence 230–231
- Goodness-of-fit statistics, for partially
classified data 326–328
- Grouped and rounded data 355–361
censored normal data with covariates
358
- exponential sample 355–356
- normal data with covariates
357–358
- Growth curve models 270
- h**
- Health and Retirement Survey
358–359
- Healy and Westmacott method 35,
266
- Heckman's model 369, 371
- Henderson likelihood 145
- Hierarchical loglinear models. *See also*
categorical data
- Historical heights 17–18
- Horvitz–Thompson estimator 24, 53,
64, 88
- Hot-deck imputation 68, 76–81, 361
within adjustment cells 78
matching metrics for 79–80
multivariate missing data 79–80
random sampling 77–78
random sampling without
replacement 78
sequential 83
single-partition vs. p-partition
79–80
- Hybrid maximization methods
208–210
- i**
- Image reconstruction 317
- Implicit imputation model 68, 75
- Improper multiple imputation
100–101, 238–240
- Imputation 67–82
comparison of methods for bivariate
data 73
for repeated-measures data 80–81
multiple, *see* Multiple imputation
hot deck, *see* Hot-deck imputation
of draws from a predictive
distribution 73–81
loss of efficiency 74
of last observation carried forward
80
of least-squares estimates 33–34
and iterating 34–35. *See also* EM
algorithm
uncertainty, estimation of 85–105
by modifying imputations 86
by multiple imputation 86
by resampling methods 86, 90–92

I
Imputation (*Continued*)

- valid methods from a single filled-in data set 86–89
- using explicit models 73–76
- using implicit models 76–80. *See also* Hot-deck imputation
- Independent variables, missing values in 267. *See also* Regression
- Inestimable parameters 38, 180
- Inestimability of models 386
- Internal Revenue Service (IRS) income data 371
- Intraclass correlation 263
- Inverse-probability weighted (IPWGE) 27, 299
- Inverse-Wishart draw 132
- Iterated conditional modes algorithm 201
- Iterative algorithms, *see* Algorithms, iterative
- IVEWARE program for multiple imputation 243

j

- Jackknife repeated replications (JRRs) method 361
- Jackknife standard errors 92–95
 - for complete data 92–93
 - for imputed data 93–95
- Jeffreys' prior distribution
 - for factored monotone normal sample 159
 - for multiple linear regression 114–115
 - for normal pattern-mixture model 373
 - for normal sample 128–129
 - for multinomial model 225–226
- Jensen's inequality 194

k

- Kalman filter models 276–277

I

- Lack of fit 244, 328, 388. *See also* Goodness-of-fit statistics
- Laplace (double exponential) distribution 148
- Large-sample likelihood theory, *see* Maximum likelihood, large-sample theory
- Last observation carried forward (LOCF) 80
- Latent variables 12. *See also* Factor analysis
- Latin square 34
- Likelihood
 - based estimation 109
 - direct vs frequentist likelihood inference 119–120
 - equation 112
 - for exponential sample 112
 - for multinomial sample 112
 - partial likelihood 118, 140
- Linear estimators 87, 94
- Louis's method for computing standard errors 220–221

m

- Mahalanobis distance 289
- MAR, *see* Missing at random (MAR)
- Markov-chain Monte Carlo (MCMC) methods, *see* Bayesian inference
- Markov model 188
- Matrix sampling 18–19
- Matching to fill in respondent values, *see* Hot deck imputation
- Maximizing likelihood over the missing data 141–145
- Maximum likelihood (ML) 141. *See also* EM algorithm; Likelihood estimate, definition 112
- estimation with complete data 109–118

- for one-to-one functions 113, 116
 for specific missing-data problems:
 ANOVA 265–266
 autoregressive time-series models 273
 censored normal data with covariates (Tobit model) 358
 contingency tables 301–326
 factor analysis 261–262
 grouped exponential sample 355–357
 grouped normal data with covariates 357–358
 growth curve models 270–273, 294–297
 ignorable 133
 linear regression 264–269, 344–346
 logistic regression 346–347
 log-linear models for contingency tables 320–325
 MANOVA 268–269
 mixed continuous and categorical data 331–333
 multinomial data 190, 313–317
 multivariate contaminated-normal data 290–291
 multivariate normal model 250–252
 multivariate linear regression 268–269
 multivariate t model 290–291
 nonignorable models 138–139, 302, 354–355
 robust regression 297. *See also* multivariate t model, contaminated normal model
 repeated-measures models 269–273, 294–297
 restricted covariance matrix 257–264
 stochastic censoring (Heckman) model 362
 time series 273–277
 univariate normal sample 189
 univariate t sample 197–198, 286–287
 variance components 262–264
 large-sample theory 119–126
 situations where ML fails 120–121
 MCAR, *see* Missing completely at random (MCAR)
 Mean and covariance matrix,
 estimation from incomplete data 25–26, 61–63, 69–70. *See also* Multivariate normal data with missing values
 bias of complete-case analysis for a mean 48–49
 robust estimation 288–291
 Mean imputation 68, 360. *See also* Filling in for missing values
 Measurement error
 as missing data 19–20
 auxiliary variables subject to 357, 75–376
 Mechanisms that lead to missing data 13–23. Method of weights 344
 Metropolis–Hastings algorithm 232
 MICE program for multiple imputation 243
 Mills ratio 358
 Missing always completely at random (MACAR) 14, 182
 Missing always at random (MAAR) 136, 182
 Missing at random (MAR) 14
 Missing completely at random (MCAR) 14–15, 19–23, 25, 28, 48–50, 54, 61–63, 69–70, 72–75, 77–78, 85, 99, 150, 156, 158, 177, 251–253, 270, 306, 313, 326, 388

- Missing data
by design 18–19
codes 4
Covariates in regression 74–75,
329
definition of 4
ignorable 132–134
in clinical trials 13
in experiments 29–45
in multiple-user data bases
101
indicator matrix 8–9, 13
literature reviews 23–24
mechanism 8, 13–23
pattern 8
taxonomy 23–28
- Missing-information principle 188.
See also EM algorithm
- Missing not at random (MNAR) 14,
16, 21, 24, 351–404
mechanisms 371
missingness mechanisms
355–361
models for categorical data 385
models for categorical data
385–390
models for repeated-measures data
382–385
nonresponse 364
parameter restrictions for
340–343, 369
- Missing-plot techniques, *see* Analysis
of variance (ANOVA)
- Missing-value covariates 35
- Missing values as parameters
141–145
- Missingness
indicators 9
known vs. unknown 11–13
mechanism 8–13
patterns 8–13
- Mixed-effects analysis of variance 30,
262
- Mixed normal and nonnormal data
329–349
- Mixture models 345–346
- Model-based procedures for missing
data 25
- Monotone pattern of missing data 11
filling in data to create a 239
for bivariate counted data 303–305
for bivariate normal data 153–161
for multivariate counted data
305–308
for multivariate normal data 26,
161–175
- Monte-Carlo simulation, *see* Bayesian
inference, simulating draws
from posterior distribution
- More observed variables, *see*
Monotone pattern of missing
data
- M step (maximization step) 188–193.
See also EM algorithm
- Multinomial data 302. *See also*
Categorical data
factored likelihoods for monotone
302–313
standard errors for 215–216
- Multiple imputation 67, 74, 95–100,
232–245, 351
approximations for P-values
235–238
approximate ways of creating
238–239
using asymptotic distribution of
parameters 239–240
using importance sampling 240
using the bootstrap 240
- Bayesian theory 232–235
- bivariate normal example 235
- compared with resampling methods
100–101
- improper 238–239
- inferences 351
- MICE 243

- misspecification of imputation model 27
 proper 100
 stratified random samples 98–100
 uncongeniality between imputation and analysis model 244
 Multiply-imputed data set, analysis of 97, 232–245
 Multiple linear regression, *see* Linear Model; Regression
 Multiple maxima of likelihood 252, 333, 342, 345, 355
 Multiply-imputed data set 97
 Multivariate analysis of variance 268–269
 restrictions in general location model 344
 robust 293
 Multivariate normal data with missing values 249–257
 Bayes for monotone missing data 174–175
 Bayes inference for general pattern by data augmentation 253–255
 estimated asymptotic covariance matrix 252–253
 estimation with restricted covariance matrix 257–264
 example using St. Louis risk research data 255–257
 ML for general pattern by EM 250–252
 ML for monotone missing data 26, 161–175
 ML for special nonmonotone patterns 175–182
 Multivariate regression 228–229
 Multivariate t distribution 31, 129, 347
 Multiway tables 236, 314, 318, 341 *see also* Categorical data
- n**
 Net worth imputation from interval-censored data 359–360
 Never jointly-observed variables 12
 file-matching problem 12–13
 Newton–Raphson algorithm 186–187
 Noncontact 9
 Missing not at random (MNAR)
 model 351
 Nonlinear regression 142
 Nonresponse, strata 5, 7
 Normal data, censored 357–358
 grouped with covariates 358–359
 linear regression model, *see* Linear Model, Regression
 nonignorable models 357
 Not missing at random (NMAR), *see* missing not at random (MNAR)
- o**
 Observed likelihood 201
 Odds ratio, ML estimate from partially classified data 327
 Outliers, *see* Robust estimation
- p**
 Pairwise available-case methods 62
 Parameter-expanded EM algorithm, *see* PX-EM algorithm
 Parameter-expanded DA algorithm, *see* PX-DA algorithm
 Partial correlation 177, 178
 Partially-classified contingency tables 301–302. *See also* Loglinear models
 Partially missing at random (P-MAR) 140
 Patterned covariance matrices 259
 Pattern-mixture models 362–365
 survey nonresponse 353–355
 univariate missingness 362–364
 univariate nonresponse 352

- Pattern-set mixture models 353–355
 Pattern of missing data 11, 12
 Pearson chi-squared statistic, definition 319. *See also* Likelihood ratio statistic
 Penalized spline of propensity models 298–300
Pivoting, *see* Sweep operator (SWP)
 Poisson model in practical situations 302
 Polling data, Slovenian Plebiscite with 8, 389–390
 Positron emission tomography (PET) 316
 Possibly incompatible Gibbs' sampler (PIGS) 353
 Posterior distribution, *see* Bayesian inference
 Posterior standard errors 221–222
 Poststratification 58–59
 Potential outcomes as missing data 5
 Power transformation 369
 Predicting missing values, *see* Filling in for missing values
 Principal component analysis 249
 Prior distribution, *see* Bayesian inference
 Probability of response, *see* Response propensity
 Proc Mixed, *see* SAS software
 Propensity scores 56, 298
 stratification on 57
 weighting by inverse of estimated 58
 Proper multiple imputation 238
 Proportional-hazards models 400
 Proxy pattern-mixture analysis (PPMA) 378
 P-step (posterior step) of data
 augmentation 224–225
 PX-DA algorithm 229
 PX-EM algorithm 206–208
 for factor analysis 261–262
 for univariate t sample with known degree of freedom 206–208
 rate of convergence 208, 216
- q**
 Q-function 204–209. *See also* EM algorithm
 Quality of life data 5–8
 Quasi-Newton acceleration method 209
 Quasirandomization inference 54
- r**
 Raking-ratio estimation 58–60
 Random-effects model 117
 Randomization inference for surveys 50–51
 Randomized block 39–40
 Randomly missing data, *see* Missing completely at random (MCAR); Missing at random (MAR)
 Random sampling with replacement 77–78, 85, 87, 90, 219–220
 Rate of convergence 35, 198–200
 Ratio estimator 149
 Refined and coarse classifications 309–312
 Refusal to answer 10
 Regression
 calibration for measurement error 75–76
 computation via SWEEP 167
 estimator 154
 imputation 68, 71. *See also* Buck's method; Filling in for missing values
 interactions in 267
 measurement-error regression 75–76
 missing covariates in 74–75
 small-sample inference of 158–161
 subsample-ignorable likelihood for 379–382
 stochastic 73

- Regular exponential family 116, 196, 210, 250, 314
- Repeated imputations 97
- Repeated-measures model, with missing data 269–273
- Replacement units, *see* Substitution of Missing Units
- Resampling methods, bootstrap and jackknife 90–92
- Residuals, added to imputations 68
- Response propensity, *see* Propensity scores
- Response rate 66. *See also* Pattern of missing data
- Restricted covariance matrix 257–264
- Restrictions on cell means in general location model 337–340
- Reverse sweep (RSW) 169–170
- Reviews of missing-data literature 47
- Right-censored survival 16–17
- Robust estimation 285–300
- adaptive 291–292
 - for univariate samples 286–288
 - inference 365
 - of means and covariance matrix, with complete data 288–289
 - of means and covariance matrix, with missing values 290–291
- Rounded data, *see* Grouped and rounded data
- S**
- Sample surveys 86
- Sampling importance resampling (SIR) 232
- Sampling weight 24, 52
- Sandwich estimator of asymptotic covariance matrix 123
- SAS PROC MI 241
- SAS software 269
- Satterthwaite approximation 98
- Saturated model for contingency table 320–325
- Score function 112
- SECM algorithm 222
- Selection model formulation 352
- SEM algorithm 214–219
- applied to ECM and PXEM iterates 222
 - monotone bivariate normal sample 216–219
 - multinomial data 215–216
- Sensitivity analysis for nonignorable nonresponse 376–378
- for chained-equation multiple imputation 241, 391
- in survival analysis 344, 355
- pharmaceutical applications 396–397
- Shared-parameter models 384–385
- Simplified selection factorization 353
- Simulation studies of missing-data methods 63
- Single imputation 67
- Slovenian Public Opinion Survey 390–391
- Software, *see* Computer software for missing data methods
- Space-filling condition for ECM convergence 201–202
- Speed of convergence, *see* EM algorithm, rate of convergence
- Speeding convergence, *see* Acceleration techniques
- S-plus software 269
- Standard errors
- based on information matrix 213–214
 - using Louis's method 220–221
 - in ANOVA 41–42
 - from cluster samples 89, 94–95

- Standard errors (*Continued*)
 of estimates, large-sample theory.
See also Asymptotic covariance
 matrix of parameters or
 estimates
 via other methods 214–219
 using Bayesian methods 221
 using the bootstrap 219–220
- Starting values for algorithms 252,
 328
- State-space models, *see* Kalman filter
 models
- Stationary time series 274–276
- Statistical packages, *see* Computer
 software for missing data
 methods.
- Stem and leaf plot 15
- St. Louis Risk Research Project
 255–257, 268–269, 332–333,
 336–337
- Stochastic-censoring model, *see*
 Censored data
- Stochastic regression imputation 68,
 73
- Stratification on propensity score
 56–57
- Stratified random sample 51, 64
- Structural zeros, contingency tables
 302
- Student's *t*-distribution, *see*
t-distribution
- Subsample-ignorable likelihood (SSIL)
 379–380
- Subsampling nonrespondents 87,
 240
- Substitution of missing units 68–69
- Supplemental EM algorithm, *see* SEM
 algorithm
- Survival data 5–8
- Sweep operator (SWP) 166–167
 applied to regression 167–169
 applied to time series with missing
 data 275–276
- t**
- Taylor series expansion 60–61, 90,
 120, 121, 374
- t* distribution 31
 ML for sample from 197–198,
 204–205
 Bayes inference for sample from
 229–230
- Time-series models 273–282
- Tipping-point analysis 355,
 400
- Tobit model 358–360
- Transformations:
 Box–Cox in stochastic censoring
 model 369–371
 of normal parameters 153
 using sweep 169–170
- Treatment discontinuation in clinical
 trials 13
- Two-way contingency table
 385
- u**
- Ultimate clusters (UC's) 86–89
- Unbalanced data in ANOVA, *see*
 Analysis
 of variance with missing outcomes in
 repeated measures data
- Unconditional mean imputation, *see*
 Imputation of means
- Uncongeniality in multiple
 imputation 244
- Uniform diffuse prior distribution
 148, 149, 221, 233
- Uniqueness matrix 261
- Unit nonresponse 10, 353–355
- Univariate missingness data 9–10,
 363
 normal 15–17
- Probit selection model for
 363
 normal pattern-mixture model
 363–364

v

- Variance-components models 262–264
Variance estimation, *see* Asymptotic covariance matrix of parameters

w

- Wald statistic 124
Weighted
complete-case analysis 50–58
generalized estimating equations 57–58
inference from 60–61
least squares 115, 198, 286
response rate 53

Weighting 24, 50–61

- class adjustments 54
class estimator of the mean 52
inference from weighted data 60–61
mean squared error of estimates 54
propensity 56
relationship with regression
imputation 71

Wishart distribution 130

- Within-imputation variance 97, 361
Without-replacement sampling 78
Woodbury’s identity 43

y

- Yates’s method 33

